The Effect of Humour in Political Messaging: An Investigation Combining Fine-Tuned Neural Language Models and Social Network Analysis

Janice Butler

Graduate School of Communication

Research Master's Thesis

Master's programme Communication Science

Supervisor: dr. Damian Trilling

Student number: 12356093

June 25, 2021

Word Count: 10,292*

* Agreed upon with supervisor

Abstract

"Remember when Trump forgot Mike Pence's name and then tried to have him killed?" Whether with irony, wit, sarcasm or benevolence, humour in politics brings things to the point, can evoke strong emotions or – as in this example – polarize like nothing else. The dearth of research on measurement of humour would seem to imply – for such a fascinating and powerful aspect of communication – that we are far from solving this comic conundrum.

By implementing an automated detection mechanism for humour using neural language models, it was nonetheless possible to conduct a nuanced humour-analysis of more than 420,000 political as well as 250,000 non-political social media posts. The resonance in social networks was correlated with the type and degree of humour of each posting to create a picture of which types and intensities of humour are most frequently and most successfully employed by politicians, political journalists and comedians.

Key findings are, that humour invariably provides an effective boost to message propagation and heightened comedic intensity maximises propagation in all groups analysed. When politicians use humour they prefer sarcasm to all other types, but message propagation succeeds for them better with ironic humour. Benevolent humour is least used, but produces the second best propagation effect for politicians. Political journalists – like politicians – prefer to use sarcastic humour, but they achieve most resonance with cynical remarks. Humour is, however, not always the ideal communication tool, is somewhat sparingly used by politicians (< 9% of all tweets) and even more so by journalists (< 4%). Of the 5% most liked postings, the majority in every group – including comedians – were non-humorous.

The automated measurement of humour-type and degree enables processing of very large amounts of text in a very short space of time. The approaches to network analysis allow an immediate impression of relative take-up by the information consumers. Whether the meaning of humorous texts is being deciphered or whether rhetorical techniques of humourists are merely being recognised is debatable. The ability to quantify reliably "the packaging" is in any case a crucial step forward in automated content analysis.

The Effect of Humour in Political Messaging: An Investigation Combining Fine-Tuned Neural Language Models and Social Network Analysis

Humour is a key factor for influencers, and on social media networks humour can be a powerful and convincing initiator of information propagation (e.g. if used by an opinion leader). In a political setting the right kind of humour – as a tool of effective communication – enthuses voters and evokes a “man of the people” effect (e.g. Martin et al., 2003; Cheng et al., 2021; Hendriks & Strick, 2019). Humour can thus be leveraged effectively if the right kind and intensity can be found, providing differentiation from other information sources and increasing viral-propagation.

The automated detection of intended humour is becoming increasingly vital. A chatbot, which doesn't “get the joke" or recognise the irony of a remark may go off at a dangerous tangent. The measurement of humour (comic style[1] and degree) would be a crucial tool needed to determine in big-data media experiments, where the application – or indeed omission – of humour (through A/B testing) can provide impetus (or a brake) to the propagation of information (Davis et al., 2018).

While some newer research promises the possibility of binary determination of humour (funny/not funny), the comic styles and the degree of humour have not, as yet, to the best of my knowledge been reliably detected. Whilst some research has been dedicated especially to sarcasm e.g. Badlani et al. (2019), natural language understanding (NLU) is still missing several pillars required for true comprehension (as opposed to the processing of mere semantics), one of which is differential understanding of humour. This research may, in this respect, help in closing one gap in decrypting communicative intent and applies various sentence embedding techniques to create an optimized “humour detector” in order to advance NLU in the field of political communication.

[1] fun, benevolent humour, nonsense, wit, irony, satire, sarcasm, and cynicism

Most current research is based on creating large annotated datasets to be used in the supervised training of neural networks e.g. Fan et al. (2020). As the technique of using Neural Language Models (NLMs) and the most recent advancements – whilst still quite young – show great promise (e.g. sentence embeddings used by Annamoradnejad & Zoghi, in 2021), the approach for this paper was to investigate and incorporate these developing techniques and, in an initial project phase, create an optimal humour detector.

Apart from the technical questions posed in the paper, the societal question investigates how humour is employed in political news i.e. by politicians and the journalists covering them. Moreover, a determination of which types of humour and which degree have the most marked effect for each group was investigated.

To address this underlying question, the principle (neural language) models under scrutiny were subjected to the successive application of two types of fine-tuning for use in the domain of humour-detection and evaluation. The training data was based on manual annotation of data acquired from multiple sources together with repurposed corpora re-annotated for this project. Verification of results was determined statistically using validation based on segmented training, evaluation and test datasets (Brownlee, 2018).

Having established the means for humour-degree and classification, the determination of the effects of humour on political messaging were then achieved through cross-referencing the humour of large numbers of political tweets with their “virality” as determined via analysis of the tweets’ effectiveness in the social network. As a control experiment, large numbers of tweets from particularly comedic sources were also measured for humorous content and compared with those of the journalists and politicians.

Finally, recommendations for optimal comedic style and levels of humour became possible, with the tools to assess prospective political messages in these regards.

Theoretical Background

In the following, the fundaments of comic styles and humour in politics are addressed; the concept of opinion leadership and its measurement is explained, as are the ideas behind NLMs and social network analysis, these being the tools used respectively for automatic humour detection and the identification of viral-propagation in social media networks.

Comic Styles

Humour emerges as a political communication device used by citizens and not only by traditional actors such as journalists or the politicians themselves. As Davis, Love and Killen (2018) explain, the new participatory medium of the internet, and "internet culture" which rewards humour, has promoted humour's incorporation in political discussions. As Hennefeld (2016) explains, in recent decades through shows like *John Oliver's Last Week Tonight* the "cultural economies of laughter have become inextricably entangled with […] civic processes". Thus, the use of humour in political communication has had a resurgence in popularity and impact.

Ruch, Heintz, Wagner and Poyer (2018) describe different types of humour in a notion they call "comic styles". Accordingly, there are eight distinct types, which can be considered at a fine-grained level. Other researchers differentiate between further comic styles. Lauer's (1974) classification includes for instance "comic in a narrow sense" and instead of mere irony, he distinguishes specifically self-irony as well but omits to mention "nonsense" at all. Milner Davis' (2003) description on the other hand goes into much greater depth mentioning "romantic/festive or sentimental comedy (e.g. sitcoms)" amongst others.

The classification according to Ruch et al. (2018) and subsequently Mendiburo-Seguel and Heintz (2020), as employed in this paper, differentiates further between darker types of humour: sarcasm, satire, cynicism and irony – identifying mockery/ridicule as its essence – as opposed to lighter types. An adaptation of the descriptions of Ruch et al. was chosen due to its comprehensiveness and the direct application to textual content (as

opposed to other media). Certain key words were picked out of the theoretical descriptions to guide the annotators in recognising the differences between styles see Table A1 in Appendix A. In the following sections, distinctions and overlaps between humour-types are discussed.

Dark Comedic Styles.

Irony. Critical statements can be found not only in irony, but also in other dark humour-types (i.e. satire, sarcasm and cynicism). In the case of irony, these statements take an evaluative turn. The incongruence between the literal meaning and the context is the identifying feature of irony. Ironic statements are appreciated by the intelligent while mocking the stupid (Ruch et al., 2018).

Cynicism. Modern day cynicism is differentiated by Higgie (2014) as knowing that people exist in a world of meaningless constructions[2]. The central identifying factor is the outlook on the world i.e. the moral values of the author. While satire also addresses the moral values and aims to improve humans, this sense of improvement is not present in cynicism. The cynic holds moral values in disdain and considers them ridiculous.

Satire. Satire is more aggressive than cynicism and aims at detecting weakness, though this is paired with attempts at goodness. Not only deprecating of the bad and foolish is involved, but also the intention of improving the world and rectifying fellow humans. Both cynicism and satire have aggressive tendencies. Mockery is for pure pleasure, but is based on a morality-based criticism (Ruch et al., 2018).

A satirist reveals the true circumstances of the world (Ruch et al., 2018). There is also a potential threat to the self-image according to research by Boukes et al. (2015) – as the reader himself may be a victim of the criticism – leading to audiences being unamused even

[2] giving up (cynicism) and instead subverting and exposing these constructions (kynicism)

though the joke may be understood and the threat to self-image may result out of a misunderstanding of the intentions of the author.

Sarcasm. Joshi, Bhattacharya and Carman (2016) write that an intent to mock must be present, be it directed towards institutions or people. The aim is to expose a corrupt world, whereas cynicism goes a step further, the mockery concentrating on moral values, which are considered ridiculous.

Light Comedic Styles.

Benevolent Humour. This type of humour has also been referred to as a defence mechanism (due to self-deprecating elements) (Freud, 1928), coping humour (Martin & Lefcourt, 1983), self-enhancing humour (Martin et al., 2003) and humour as a character strength (Peterson & Seligman, 2004). The goal is always to foster understanding and sympathy with others (Ruch et al., 2018).

Fun. Fun, like benevolent humour, also shows social elements, though in this style teasing or mischievousness may be used, whereas benevolent humour is generally accepting of mistakes. Banter is often used to make a point, to defeat boredom and for socialization purposes (Plester & Sayers, 2007).

Nonsense. While ridiculous aspects of things are presented, there is no ulterior motive displayed, as opposed to benevolent humour and fun, which aim to create a good mood. Logic and incongruity are used to play with language. An upside-down world is created through bizarre or fantastical stories (Ruch et al., 2018).

Wit. Wit intends to enlighten through surprising punchlines using unusual combinations and quick, appropriate remarks. In order to maximize the “funniness” the “victims” may not be shown sympathy (as opposed to benevolent humour). Links between ideas or thoughts that are not necessarily connected are made through reading the situation.

Viral Political Humour

Twitter especially is used as a space to create and share political opinions through tweeting, liking and retweeting. Mukhongo (2020) describes how the usage of humour helps to express serious political topics and that the use of tweets and memes could be interpreted as a form of civic engagement. She concludes that while it is easy to dismiss memes and other humorous content as mere amusement, the participatory culture which results from creating and sharing, does play a significant role in driving political contestation and is able to mobilise collective action as analysed in the 2017 Kenyan election. The article builds on Jenkins, Ford and Green's (2013) book exploring humour as a form of activism. Jenkins et al. (2013) pronounce that media now has to be "spreadable"[3]. Spreadable media according to Jenkins et al (2013) is classified according to seven types of content, one of which is humorous content. The consumers view content through their own experiences and ultimately respond to it.

Kopper (2020) points out that humour is dependent on culture and context, which means that some may not understand the joke conveyed. Additionally, while something may have been funny in the past, the context might have changed, making the joke inappropriate or no longer funny. If individual (distinct from contextual) differences matter whether a statement is interpreted as being funny or not, it is also possible that it matters what type of humour is employed. Some – or by extension, groups of – people might, for instance, prefer (or understand) sarcasm more than nonsense humour.

Opinion Leadership

As humour is often used for the purposes of message propagation and differences in preferences of types of humour may influence the reception of a message, it is possible that

[3] Although with Web 2.0 the consumers of the content have become able to add and respond to the content created, all media are not created equal and some have a higher perceived value.

some types of humour fare better than others. The identity of the author of a humorous statement sometimes imparts additional influence on how far and wide the message is spread, if only due to that person's general higher visibility and followership.

Lazarsfeld, Berelson and Gaudet in their study "the people's choice" (1944) first explained the concept of opinion leaders, who are assumed to exert higher influence[4] on individuals in their social circle and can thus be considered as role models. The development of social media networks aids in the potential influence an opinion leader may have, since communication in social media involves a potentially much larger audience. However, the targeting of a specific audience is limited to cases of publicized reactions such as through a change in status, tweeting or announcements on Facebook, as explained by Winter and Neubaum (2016). Moldovan et al. (2017) expand that opinion leaders are highly effective in a close, small, social network but that their influence declines in larger groups with weaker ties.

Virality

There are several ways to identify opinion leaders, as Moldovan et al. (2017) elaborate. One would be the use of self-reporting scales or – more reliably – by asking others to identify opinion leaders. However, as this paper involves conducting an automated analysis of leading opinions, metrics measuring importance (of a politicized message) within a network involving large amounts of data are required. In network analysis the nodes represent individuals, information instances, organisations or other social entities. Nodes are connected through edges representing relationships, which could be friendship, an exchange of information, references (such as retweeting on Twitter) or sharing (Xu et al., 2014).

There are very many different types of analysis that can be performed on a social network. Riquelme and González-Cantergiani (2016) have provided an extensive overview

[4] The hypothesis was that the influence of personal connections on voting behaviour is more important than that of the media.

regarding the analyses, which can usefully be performed with regard to the available data. Relationships analysed in centralities (such as betweenness or degree centrality) primarily concern those between users. However, for the purposes of this paper the influence measures of individual tweets are the focus, i.e. the numbers/extent of retweets, likes and replies.

Riquelme and González-Cantergiani (2016) also provide a list of metrics that are relevant for the analysis. In this paper exclusively original tweets (referred to as OTs) were analysed. Through the Twitter API several metrics were collected that are readily available and of relevance (explained in the Method section).

An analysis of resulting user-to-user interactions is out of scope. Analyses such as ReachBuzzRank[5] and ACQR[6] are necessarily restricted to a particular topic. Such approaches were initially considered for this project, but proved too restrictive in terms of capturing sufficient and balanced numbers of messages from the target actors. Instead the approach of defining large, representative groups of relevant actors (specifically, all UK MPs on Twitter and UK political journalists) and harvesting all tweets in a particular (2 year) period was chosen[7]. The choice of Twitter handles are elaborated upon in the Method section.

Automating Humour Detection

In order to automatically detect what comic style is used in a text and how funny (or not) the text is, that text first needs to be represented in a meaningful numeric format. Word embeddings, the basis for NLMs, are vector representations of words and their relationships. Semantically similar words are closest to each other and due to the high-degree of the vectors (tensors) used, very many relationships are represented within the embedding for

[5] predicting influential users through an analysis based on the PageRank algorithm and a temporal analysis
[6] used for detecting activity and the reputation of users
[7] Another informative technique which could be combined with this humour analysis would be (for a specific topic) the Time Network Influence Model (TNIM) to estimate effects in relation to the time intervals between messages – relevant if a specific political event were to be analysed longitudinally.

each and every word. NLMs are even more sophisticated versions of word embeddings where the word order and context – due to the pre-training involving vast amounts of meaningful texts – is taken into account. Previous approaches to humour measurement (Joshi et al., 2016) were typically limited to a single comedic type and (in some cases) its polarity or focused just on the binary detection of general humour (Annamoradnejad & Zoghi, 2021). Given that words, sentences and entire paragraphs can be finely categorised by neural models the question is posed:

> **RQ1**: Is it possible to automatically detect different degrees and/or types of humour in political messaging using fine-tuned NLMs?

Neural Language Models

Two main types of language model are compared, the AR models[8] and the AE models[9]. To understand how these types of models work, and to understand that the results are not attained by something akin to black-magic, there are two key concepts – Attention and Transformers – which need to be understood.

The Attention Model. Attention can be seen as analogous to (the technique of) extracting a specific piece of information from a textbook. Reading the whole book would certainly be an option, but is very costly of our time. Alternatively, a search for the topic in the index allows a jump directly to the appropriate pages. So, the attention is guided to only the most important areas.

The attention model (conceived by Bahdanau et al, 2014) originally addressed challenges in machine translation, but has been more widely applied in the field of AI and to other language model-based tasks, such as categorization. The weight (or attention) applied

[8] examples of which would be XLNet, GPT and its successors GPT-2 and GPT-3
[9] such as BERT and adaptations like RoBERTa or DeBERTa

to each word (token) is used in predicting what the next word should be, based on the importance of the following input token.

Why is this necessary? Long sentences cannot be fluently translated by merely proceeding word-for-word, because in different languages the word order and even the number of words, needed to express certain concepts, differs greatly. A human translator takes into account the whole sentence and sometimes the wider context of the document to understand and interpret the meaning. For machine understanding the problem is the same – a wider context than a single word or even bi- and tri-grams is necessary – with the consequence being a requirement for some form of longer-term memory.

Applying a computationally very efficient vector-based similarity measure[10] of the preceding words in a sentence ensures that, of all the previous words, only the most probably relevant are used in predicting the next one. This mechanism, called masked-attention, facilitates long-term memory at little cost, since in this way only a fraction of the preceding text (i.e. the relevant parts) is used in decision-making (Vaswani et al, 2017). Put succinctly: attention is content-based querying.

A further refinement of the attention mechanism is multi-head attention, by which parallel processing of different positions yields multiple outputs, which are then concatenated. This approach, effectively implementing several attention layers in parallel, was found by Vaswani et al. (2017) to add stability and robustness to the architecture. Long-range attention is the mechanism that enables “general knowledge” which has been – to a

[10]The softmax function applies the standard exponential function to each element of an input (word) vector and normalizes by dividing by their sum. The resulting components then equate to probabilities, since the sum of all the results equals 1.

certain extent – “learned” during the pre-training of the language model on vast corpora, to be utilized[11].

Transformers. The Transformer model is an encoder-decoder framework that “relies entirely on an attention mechanism to draw global dependencies between input and output” (Vasawi et al., 2017). This sequence-to-sequence model encodes the input word-stream, encapsulating information for all input elements in the encoder vector. The decoder, with the help of the information held in the language model, converts this vector back to intelligible language. An important advantage of this architecture is that the output sequence may differ in length from the input sequence.

Two types of language models were considered. Autoregressive (AR) models are pre-trained on predicting future values from past values. They are parallel to the decoder of the original Transformer model and a mask is used on top of the full sentence so that the attention heads can only process what precedes and not what follows. This type of model is most used in text generation.

Autoencoding (AE) models, on the other hand, are employed for feature selection and extraction and as such are commonly used for sentence or token classification. They are pre-trained by corrupting the input tokens in some way and trying to reconstruct the original sentence. Those models usually build a bidirectional representation of the whole sentence (The Hugging Face Team, n.d.).

Denoising. When the data is insufficient and there are more nodes in hidden layers than there are inputs, the model encounters the problem of “learning” the network too comprehensively, leading, paradoxically, to being unable to “learn” sufficiently well when new

[11] A distinction between general attention and self-attention is drawn when the attention mechanism is applied to multiple components such as input and output (general) or within a single component (self-attention) such as a neural network layer (Lihala, 2019)

data, previously unknown to the network. Tasks like dimension reduction or representation learning will then not be performed well enough. The subtype of denoising autoencoders resolve this problem through artificially introducing noise by randomly transforming some input values to zero (The Hugging Face Team, n.d.).

Fine-Tuning of Neural Language Models. NLMs such as BERT[12] and GPT-x are general-purpose models, which have been pre-trained over a period of weeks and even months on very large, non-specialized text-corpora. These language models thus contain huge amounts of semantic information but also substantial levels of contextual information (generated sentence embeddings) regarding their vast vocabularies. However, to prepare a language model for specialized use – as we wish to do in humour detection and classification – a phase of transfer learning is required to fine-tune the pre-trained model (Howard & Ruder, 2018). Fine-tuning is comparatively inexpensive in comparison with pre-training and is the activity, which makes the NLMs practically usable in target domains such as the topic of this paper.

The differences in NLMs in tackling representations of language lead to the following research question regarding humour detection and the corresponding hypothesis:

> **RQ2:** Are there differences in detection-precision of the degree or type of humour using language models based on the autoregressive pattern (e.g. GPT-2/3, XLNet) as opposed to those using the denoising autoencoder pattern (e.g. BERT, DeBERTa) when an equivalent fine-tuning is applied?

According to the current (10.06.2021) SuperGLUE Benchmark[13] leaderboard (apart from T5, which is a sequence-to-sequence model) it is AE models which consistently score

---

[12] BERT (Devlin et al., 2019), for example is trained on two tasks, a) to predict missing words in sentences b) to predict whether one sentence follows (is related to) another.

[13] SuperGLUE (https://super.gluebenchmark.com/leaderboard): benchmark test for evaluating general-purpose language understanding systems

higher than the first AR model on the list (GPT-3) which is in 16th place (Wang et al., 2020). This leads to the hypothesis:

> **H1:** NLMs based on the denoising autoencoder pattern are more precise in detecting degree and type of humour.

Applying Humour Detection to Politics. Based on the metrics established by Riquelme and González-Cantergiani (2016) it is thought that the measures described are ideal for an approach centred on tweets (rather than users) and are used extensively in this project. A numerical analysis of messaging at scale in social networks, offers the opportunity to gather much insight into reactions to politically inspired humour, with a view to answering the following question and testing two hypotheses:

> **RQ3:** Is there a particular type or level of humour that is most effective in achieving high levels of propagation of political messaging?

Vance argues that “in certain situations verbal irony is a more effective and thus more logical mode of speech than its literal equivalent” (Vance, 2012, p.7). Additionally, Knoblock (2016) argues that ironic statements may be more memorable and bring the speaker social recognition due to the effort invested. Certain benefits of using irony for a political actor are thus recognised and it is hypothesized that:

> **H2:** The most effective type of humour for political messaging is irony

While previously argued that the use of humour may be beneficial for propagation, this may not always be appropriate and serious messages may also achieve significant virality given the right circumstances.

> **H3:** More intense levels of humour (from a given actor) improve the propagation of political messaging, whilst entirely serious messages may still be quite effective propagators, but in proportionately fewer instances.

Method

To establish the effect of humour on political messaging in social networks, in answering the RQ's and testing the hypotheses, the project was necessarily split into two phases:

a) Development of a generalized humour-detection software (HDS) for automatic assessment of both humour-type and degree. The technique was based on the fine-tuning of pre-trained, general purpose NLMs using large numbers of annotated humorous and non-humorous texts. A wide variety of NLMs with differing architectures and sizes were fine-tuned on the annotated data, with the statistically best models (according to F1) being chosen for use in the next phase and to provide answers to RQ1 and RQ2, thereby testing H1.
b) Large numbers of tweets from heterogeneous groups of politicians, political journalists and (as a control group) comedians and satirical sources were taken, subjected to HDS-analysis and, in combination with the metadata from each tweet, provided answers to RQ3 and tested H2 and H3. To test the effectiveness of the humour-types and degrees, several one-way and two-way ANOVA's were performed.

Data Procurement

Three types of data had to be collected in addition to one derivate that was used for control purposes. The first requirement was for humorous texts, which would be annotated for fine-tuning the NLMs. Secondly, an equal number of non-humorous texts was required – also for fine-tuning and easily annotated as humour-level "none" and of type "serious". Thirdly, there was a need for large numbers of social media messages including their metadata, for the analysis of political social media propagation in respect to amusing content. In all cases, data was drawn from exclusively English-language sources.

Humorous Texts

These were deliberately gathered from multiple sources in order to capture the spectrum of different comedic styles needed for the later identification of these types. Firstly,

Reddit[14] was drawn upon, with data-selection from selected subreddits[15]. Secondly, a comedic, interactive TV show named @midnight was an abundant source of humorous tweets and finally – to ensure representation of highly rated comedic content – a compilation of the best jokes from prominent comedians completed the corpus of comedic data.

Reddit was considered a primary source, since not only could very large numbers of ostensibly specific humour-types be obtained, but, due to Reddit's online scoring system, it was also possible to directly derive a rating of the popularity and – by association with the subreddit topic per channel – the degree of humour of each text.

Non-Humorous Texts

Serious news outlets were targeted for this category of data. These were procured via the Twitter API using the following Twitter-handles: @AP, @BBCworld, @ITN, @ITVnews, @SkyNewsPolitics, @TheEconomist.

Political Texts

This category of data was drawn entirely from Twitter, since here the metadata was also required in all cases. A list of Twitter handles taken from a curated Twitter list of UK political journalists[16] (N=232) provided the source for journalistic data collection. To acquire a balanced corpus of political messages, it was decided to use the Twitter handles of all UK Members of Parliament (MPs) maintaining a presence on Twitter (N=588)[17].

Social Network Control Group

The aforementioned sources could all contain potentially humorous texts, but it was expected that these actors would produce predominantly serious texts. Thus, it was decided

[14] Reddit.com is self-described as: "a social news aggregation, web content rating, and discussion website"
[15] r/satire, r/showerthoughts, r/sarcasm and r/surreal
[16] @mousetrapmedia/uk political journalists / Twitter
[17] A current list of UK MP Twitter handles plus other statistics is available at https://www.politics-social.com/.

to add another set of texts as a control group where the explicit intention is to create overwhelmingly humorous texts. A curated list of British comedians (N=92) was added (Buzzberry, 2020) and the query time-period extended to provide sufficient data. The exact Twitter handles for all groups, can be found in Appendix F and on Github (https://github.com/jb-diplom/humour-detection/tree/main/scraping/twitter).

Data Collection Techniques

The data was obtained using two main techniques. Twitter data was collected using the Python library snscrape (JustAnotherArchivist, 2020), through which searches were submitted to the Twitter API. The queries, built from the Twitter handles and the specified time period, posed only a slight complication since each query had to be split into much smaller batches because a single query encompassing up to 588 handles (for all the UK MPs) would not be accepted by the API.

Once the data was captured, all required metadata could be saved, together with the user-Id, tweet-Id and the content itself. The metadata metrics used – employing the nomenclature of Riquelme and González-Cantergiani (2016) – are:

- M3: the number of mentions to the author by other users
- RT3: the number of users who have retweeted author's tweets
- FT3: the number of users that have marked author's tweets as favourite (likes)
- and RP3: the number of users who have replied to the author's tweets

The non-humorous corpus was scraped in the same way, except that the metadata was not saved, since superfluous to requirements.

The subreddit data were collected using the Reddit-API[18]. The method of collection was modified from the techniques of Orion Weller and Kevin Seppi ('WaterCooler: Scraping an Entire Subreddit (/r/2007scape)', 2019), which describe and provide basic Python code to scrape a subreddit for a specified time period. The data was collected in 2 passes, whereby initially the identifiers for the specified time-period were collected in a simple json format. Based on the identifiers, the additional submission-data is collected in the second pass including the up- and down-counts, from which the humour-score is calculated. Data-cleaning involved eliminating duplicate entries, excluding references to media (such as memes) and skipping shorter texts of less than 40 characters, since these proved to result in much noise.

The @midnight dataset had already been scraped by Potash et al. (2017) and was downloaded from the SemEval-2017 task (BhMadStudio, 2017). The data collected was then cleaned for further processing and manually annotated.

Utilizing Neural Language Models

All NLM processing was conducted based on the HuggingFace[19] Transformers library, which interfaces to both the PyTorch [20] (used throughout) and TensorFlow[21] frameworks. The HuggingFace repository (https://huggingface.co/models) was the source for all the pre-trained NLMs employed in this project.

The fairly homogenous set of hyperparameters across the models (i.e. aspects of the NLMs that can be tuned to optimize training) made a comparison of results possible as well as facilitating an almost identical approach to training each of the models.

---

[18] https://pushshift.io/
[19] https://huggingface.co
[20] https://pytorch.org/ an open source machine learning framework from Facebook
[21] https://www.tensorflow.org/ an open source machine learning framework from Google

Data produced during and resulting from the fine-tuning was relayed to another cloud service supporting machine learning practitioners at Weights & Biases[22]. This enabled the training progress to be tracked in real-time and served as a database to collect all results, as a repository for the completed NLMs and provided graphical representations of the results. This proved key in organising and assessing the many training-runs and in objectively deciding which combination of hyperparameters were to prove most effective. In particular the visualizations in the parallel coordinates chart (e.g. figure 1) enable an interactive, relational comparison between hyperparameter choices and output metrics (Accuracy, Precision and F1). This visual comparison made it possible to quickly home in on the best combination of hyperparameters.

The entire project was programmed using Python with Google Colab-Pro[23] providing access to GPU processing. Despite some restrictions which Colab has, such as the limit of a 24-hour run-time, restrictive access to GPUs and a maximum of two processes running in parallel, the platform performed well and can be recommended for other researchers.

For the humour-degree task the Reddit data was used. Specifically the online scores (from up- and down-votes) were proportionally segmented to fit the required scale of 1 to 5 indicating various degrees of humour, with 0 indicating no humour / serious texts and being assumed for all the texts acquired for the non-humorous dataset. Additionally, the @midnight data and the jokes from prominent comedians were manually annotated for degree of humour.

To verify the quality of the data used for the models on a subset of the annotated data the inter-coder reliability was measured through a Krippendorff alpha calculation. Conventionally, a value of above 0.667 is considered acceptable (Krippendorff, 2004)

---

[22] W&B: https://wandb.ai/site
[23] Google Colab: https://colab.research.google.com/notebooks/intro.ipynb

however, as Carletta (1996) writes: “Taking just the two-coder case, the amount of agreement we would expect coders to reach by chance depends on the number and relative proportions of the categories used by the coders” (Carletta, 1996, p.250). Thus, due to the choice of nine different categories (8 comic styles and “serious”) a lower Krippendorff alpha was expected.

Annotation. Part of the humour-type annotation task was outsourced to MTurk[24] where subsequently the inter-coder reliability between the worker and the author was found to be very low indeed (Kalpha = .005). These annotations were ultimately discarded and the task instead given to another coder where the inter-coder reliability was also initially found to be low (Kalpha = .228). After a discussion of disagreements a reasonably acceptable level was reached (Kalpha = .479) given the circumstance that there are nine categories and that the data to be annotated often contained phrases arguably comprised of multiple humour-types.

Analysis of the coding discrepancies showed that “fun” and “wit” were the pair most often confused and would have brought most benefit from collapsing into 1 category (resulting in Kalpha = 0.542). The dialogue revealed that creating an unequivocal choice is often extremely challenging (also due to some overlap in the definitions) and it was decided to continue without collapsing categories knowing that when fine-tuning is concluded, categories can still be collapsed at the inference stage.

NLM Fine-Tuning. Each annotated dataset was randomly shuffled together with an equivalent number of the non-humorous texts and partitioned[25] into 3 datasets for a) training b) evaluation and c) a holdout set for testing. The fine-tuning of the models was performed

[24] Amazon Mechanical Turk (mturk.com): a crowdsourcing platform to outsource human intelligence tasks to a distributed workforce

[25] A randomized approach using scikit (https://scikit-learn.org/)

using the Huggingface Transformers library[26] according to the documented examples for the text-classification task[27].

Testing. To evaluate the performance of a finished, fine-tuned model, various metrics can be computed, the relevant ones here being F1, Precision, Recall and ROC (i.e. receiver operating characteristic), Matthews correlation coefficient (MCC) as well as visualization in a confusion matrix. The basis for the calculation of all these metrics was to apply the holdout (i.e. test) dataset to the new model using the Transformers pipeline for inference[28]. For each text in the test dataset, the computed value for humour (either degree or type, depending on the model) was compared with the expected value from the annotation.

Precision indicates the ratio of correctly predicted positive observations compared to the total predicted positive observations (Joshi, 2016). In other words, high Precision indicates a low false positive (FP) rate. Recall is the ratio of accurately predicted positive observations compared to all observations in the same category. Good Recall equates to a low rate of false negatives (FNs). F1 indicates the weighted average of Precision and Recall and therefore takes into account both FPs and FNs (Joshi, 2016). The ROC is a graphical representation of the true positive rate against the FP-rate. The confusion matrix can be used to summarise the performance of a classifier and shows both the accurate and predicted cases per label (Joshi, 2016). For ease of interpretation, the results in the confusion matrix can be normalized to account for different proportions of each label in the test dataset and are reported here in this form.

Hyperparameter Tuning. To establish which combination of hyperparameters worked best, many fine-tuning runs were carried out, varying especially the chosen optimizer

---

[26] https://huggingface.co/transformers/training.html
[27] A modified version of the https://github.com/huggingface/transformers/blob/master/examples/pytorch/text-classification/run_glue.py script was used
[28] Huggingbase pipelines: https://huggingface.co/transformers/v3.0.2/main_classes/pipelines.html

(AdamW or Adafactor), the initial learning-rate, the warm-up period and learning decay type, number of training epochs, batch-size (which is inherently limited by available RAM and model-size), as well as the fundamental NLM-type chosen and the size of model, in cases where different sizes were available (see Appendix G for further recommendations).

Correlation with Virality

After the fine-tuning of the models the "best" according to the performance-parameters for each task was chosen to perform inference on the political texts to enable analysis regarding which humour-type is most effective in political messaging and which degree of humour is most successful in message propagation.

The Tweets collected from the 3 groups of actors were run through the HDS. These results were then numerically aggregated by means of a pivot table transformation. Each humour metric (type and degree) was then compared for effectiveness in viral propagation according to the Twitter metrics (i.e. the propagation-types: retweets, likes, replies and mentions), the data being normalized (to mean frequency) for each propagation-type and humour metric. The numbers and proportions of tweets for each humour-type and degree for each group were also determined using a pivot table for analysis.

Additional exploratory data analysis was implemented (in a Colab Notebook) by means of a facetted scatterplot to evaluate individual tweets (the texts, and humour-metrics) for a selection of the most prolific authors. A stacked histogram indicating which authors were numerically most active was also produced, the stacking separating the proportions of humour metrics for each individual.

Results

The results are ordered according to the 2 phases of the project, beginning with the fine-tuning of the NLMs and annotation of the required training data. An appraisal of the test results achieved with the various models and the choice of models for the 2 tasks follows, concluding with the analysis of the political messaging with respect to humour.

Choice of Model and Hyperparameters

A total of 68 fine-tunings were made across all tasks based on 17 different, pre-trained NLMs. A detailed documentation of the hyperparameter choices and training results is provided in Appendix B, organised according to task. An even more comprehensive set of results is available in the W&B Project Dashboard (wandb.ai) and in the repository on Github (https://wandb.ai/jb-diplom/humour-detection-models), together with the source code, training and test data used. Initially, fine-tuning for 2 distinct tasks was carried out, producing rather different results in each case.

Humour-Degree (k=6)

Training with data annotated with 5 degrees of humour plus the non-humorous category (k=6) proved to be by far the most challenging task. From the conducted training-runs (N=30) a weighted F1 value of above 0.563 was not achievable. For all models tested *excluding* the AR architectures the Precision, Recall and F1 metrics were very uniform (SD = 0.010, table 1). The AR architectures (XLNet, GPT2 and DistilGPT2) consistently achieved poorer values, especially bad being the GPT-derivates.

The results of the *AE architectures* showed very balanced values for Precision and Recall and produced these results uniformly, practically regardless of the hyperparameters chosen. In particular, variations in number of training epochs (between 3 and 50) had a surprisingly limited effect. The Matthews correlation coefficients (MCC) ranging from 0.385 to 0.429 support this impression.

Choice of Model. The marginally best model according to the F1-metric was the Electra-Large-Descriminator trained for 5 epochs. The visualization in figure 1, with this fine-tuned model highlighted, shows the complexity of combinations of (some of the) hyperparameters and how these contribute to the eventual F1 value. As can be seen, most models achieve an F1 > 0.52. Only the GPT architectures drop below 0.35.

Table 1

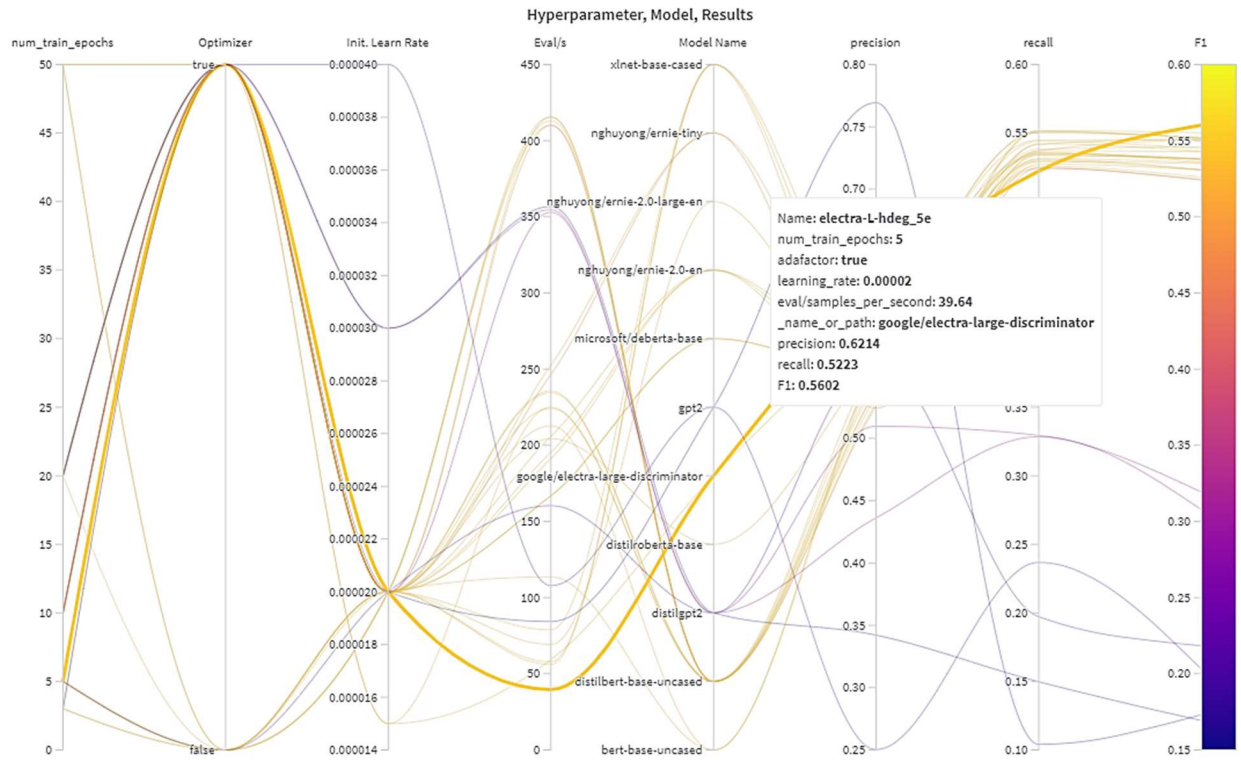
Mean precision metrics of AE and AR model architectures for humour-degree task (k=6)

Architectures	N	$M_{Precision}$	$SD_{Precision}$	M_{Recall}	SD_{Recall}	M_{F1}	SD_{F1}
All	30	0.539	0.082	0.474	0.131	0.479	0.127
AE	21	0.552	0.020	0.535	0.009	0.540	0.010
AR	9	0.463	0.141	0.320	0.165	0.326	0.157

The normalized confusion matrix for the Electra-Large was generated (figure 2). The predominance of marginal values in the upper right quadrant (excepting the serious category) demonstrates the weakness of this model, namely the presence of many FPs. Unfortunately the imprecision was even greater in the case of all other models on this task. The MCC value of 0.385, together with the ROC and Precision/Recall graphs (also figure 2) would indicate usability but only a moderate level of correlation.

Figure 1

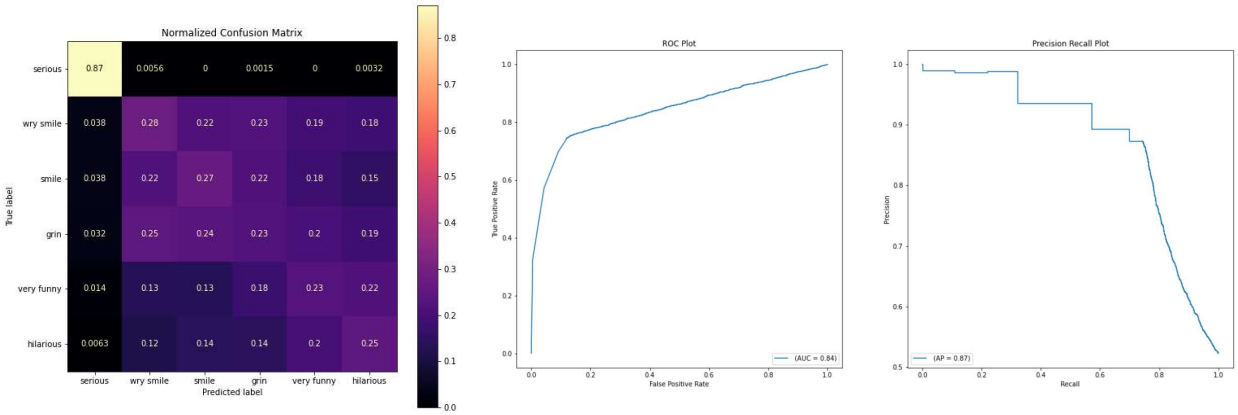
Hyperparameter Dependencies: Electra-Large (5 training epochs)



Note: Interactive plot available at: [Project Dashboard (wandb.ai)]

Figure 2

Normalized Confusion Matrix, ROC and precision recall graphs of Electra-Large (5 epochs)



Several attempts were made to improve on the results of the Electra-Large-Descriminator, by training for more epochs (10) or less (3) with differing, but negative results in both cases. The loss curves (figure G2 in Appendix G) imply that with only 3 epochs (red and purple curves), the model runs into what is known as catastrophic forgetting (McCloskey & Cohen, 1989). The training was repeated using both available optimizers (AdamW and Adafactor). With 10 epochs (green and blue curves) there is less convergence, despite considerably longer training.

Binary Humour-Degree

Due to the only moderate values achieved with k=6, it was decided to experiment with collapsing all degrees of humour into a single humour category to discover if a binary test would yield better results. These additional results are reported in Appendix C and concur well with similar approaches by Annamoradnejad and Zoghi (2021) and Weller and Seppi (2019), an F1 value of 0.997 (MCC ~ 0.993) being achieved with several models.

Humour-type (k=9)

Notwithstanding the difficulties of inter-rater reliability, the results of the fine-tunings for the 9 humour-types are very encouraging, displaying strong correlations to the test dataset for all models except the GPT2-derivates. Table 2 shows that of the conducted training-runs (N=30) a weighted F1 value of above 0.79 was achievable for all models tested *excluding* the AR architectures. The Precision, Recall and F1 metrics were very uniform for

Table 2

Mean precision metrics of AE and AR model architectures for humour-type task (k=9)

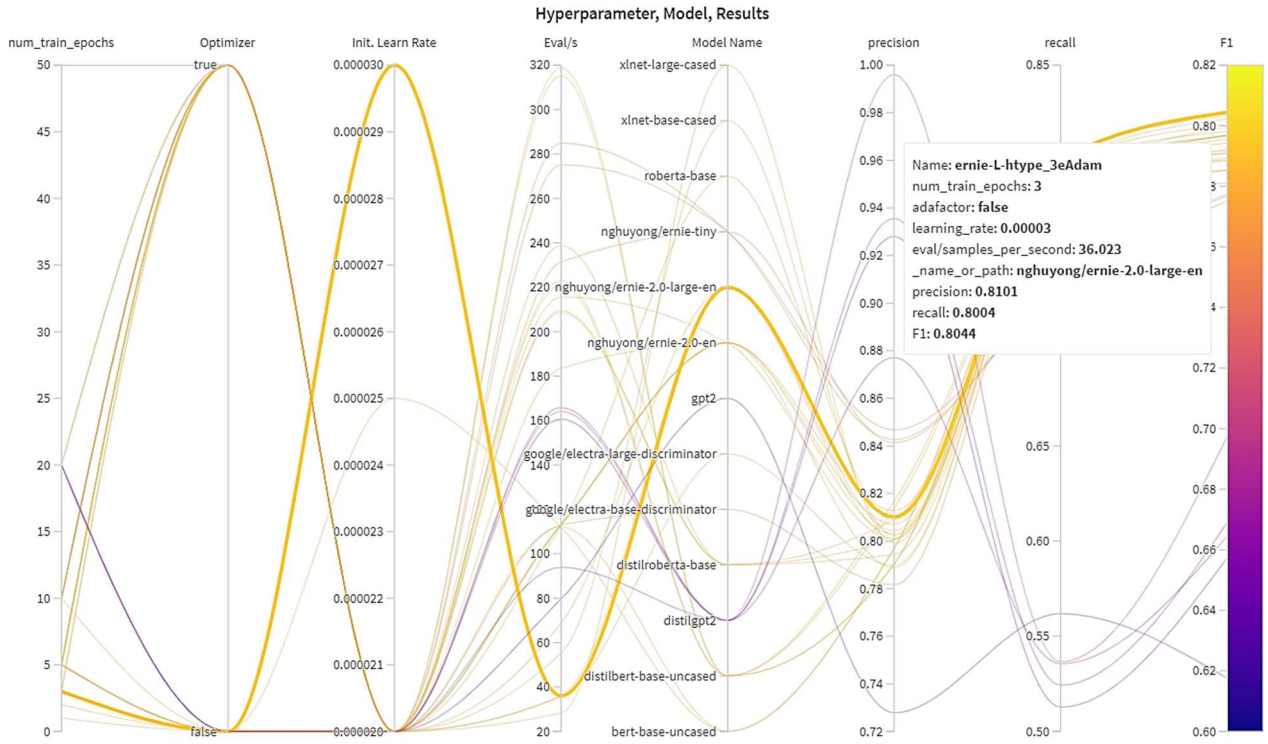
Architectures	N	$M_{Precision}$	$SD_{Precision}$	M_{Recall}	SD_{Recall}	M_{F1}	SD_{F1}
All	30	0.820	0.051	0.742	0.095	0.770	0.050
AE	23	0.806	0.017	0.783	0.019	0.792	0.008
AR	7	0.867	0.087	0.607	0.116	0.699	0.063

AE architectures. The AR architectures achieved consistently poorer values for F1 and Recall, especially so for GPT2 and DistilGPT2.

Choice of Model. The marginally best model according to the F1-metric was the Ernie-Large trained for 3 epochs (MCC = 0.701). The visualization in figure 3, with this fine-tuned model highlighted, shows the large group of models with a higher F1 and the GPT2 group below F1=0.70. The chosen "best" model was trained with the AdamW optimizer, but it is thought that this had no particular bearing, since many other models in the top group (including other Ernie-L's) were successfully trained with the Adafactor optimizer. Also noticeable in figure 3 is that the inference performance will be very poor (< 40 evaluations/s). The evaluation (and also train) speed seems largely to be reduced by increased size of model (number of hidden layers and heads).

Figure 3

Hyperparameter dependencies Ernie-L

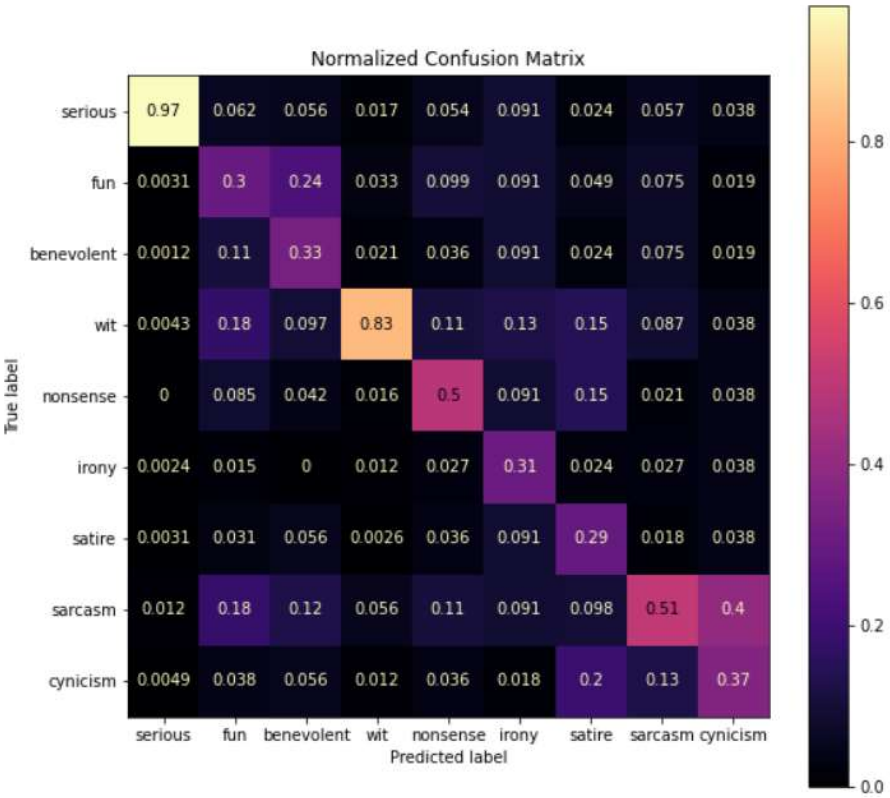


Note: Interactive plot available at: [Project Dashboard (wandb.ai)]

The normalized confusion matrix for the Ernie-Large is shown in figure 4. The clear diagonal of dominant values is most encouraging. It is only in the area of sarcasm/cynicism where much confusion arises, with FN's of sarcasm over cynicism. It should be noted that the confusion matrices of the other leading models were of a similar quality, some of which avoided (although barely) the slight majority of the sarcasm/cynicism FN. This offers potential for other techniques, for instance by combining results from multiple NLMs using an N-model redundancy approach, ensembling methods (Xu,Barth and Solis, 2019) or multi-class Boosting (Huang, She, Zhang, 2020).

Figure 4

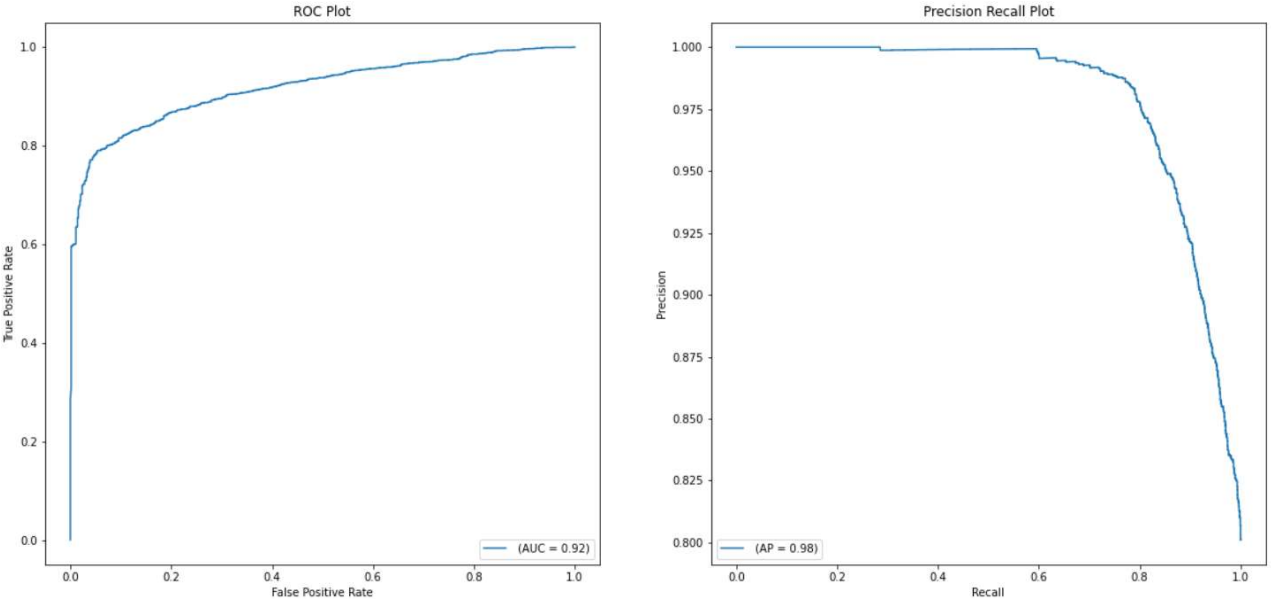
Confusion Matrix for Humour-Type Detection, Ernie-large



The ROC and Precision/Recall plots in figure 5 support the impression suggested by the confusion matrix, that a strong correlation with the test dataset was achieved.

Figure 5

ROC and Precision/Recall, Ernie-Large



Correlation of Humour in Political Tweets to Virality Metrics

For each task the best fine-tuned model (according to F1-metric) was chosen to evaluate political Twitter messages, combining with a rudimentary network analysis to determine what effect humour has on the social media propagation.

Politicians

The Tweets (N=225,100) of 588 UK MPs (> 90% of lower house parliamentarians) were collected for a 2 year period. The mean propagation results for humour-type are summarized in figure 6 and show that in all cases there are significantly more likes than retweets, more retweets than replies and more replies than quotes

To test H2 and H3 for the dependent variable, the 4 measures of virality (retweets, likes, replies and mentions) were dimension-reduced to one factor (V_{pca}) by means of a Principal Component Analysis with an Eigenvalue of 71% confirming validity of this approach. The items showed a moderate to high correlation of above .48. The Kaiser –Meyer - Olkin Measure of .66 indicated an adequate (>.5) sample size, whilst Bartlett's test of Sphericity confirmed the significance of the analysis.

Humour-Type. To test H2, two one-way ANOVA's were performed, assessing the influence of humour-type on virality for each actor (politician vs journalist). For the politicians a statistically significant effect of the humour-type was found regarding the propagation $F(8, 225059) = 60.60$, $p < .001$, Eta-squared = .002. A Bonferroni post-hoc test indicated several significant differences between each of the humour-types. The greatest impact was indicated for irony ($M = .59$, $SD = 2.68$, $n = 337$), the most marked difference being between serious and irony ($M_{difference} = -.53$, $p < .001$).

Of note: the assumption of equal variances in the population was violated for both journalists, Levene's $F(9, 194816) = 81.08$, $p < .001$ and politicians, Levene's $F(8, 225,059) = 104.33$, $p < .001$.

Humour-Degree. To assess H3, the analysis was conducted in two steps:

A1) excluding the serious category, a two-way ANOVA was conducted with V_{pca} (DV) and the types of humour (ordinal IV) together with type of actor (politician vs journalist) as the second IV.

A2) a second two-way ANOVA was performed with V_{pca} (DV) and a dummy coded IV serious/not serious and the type of actor (IV).

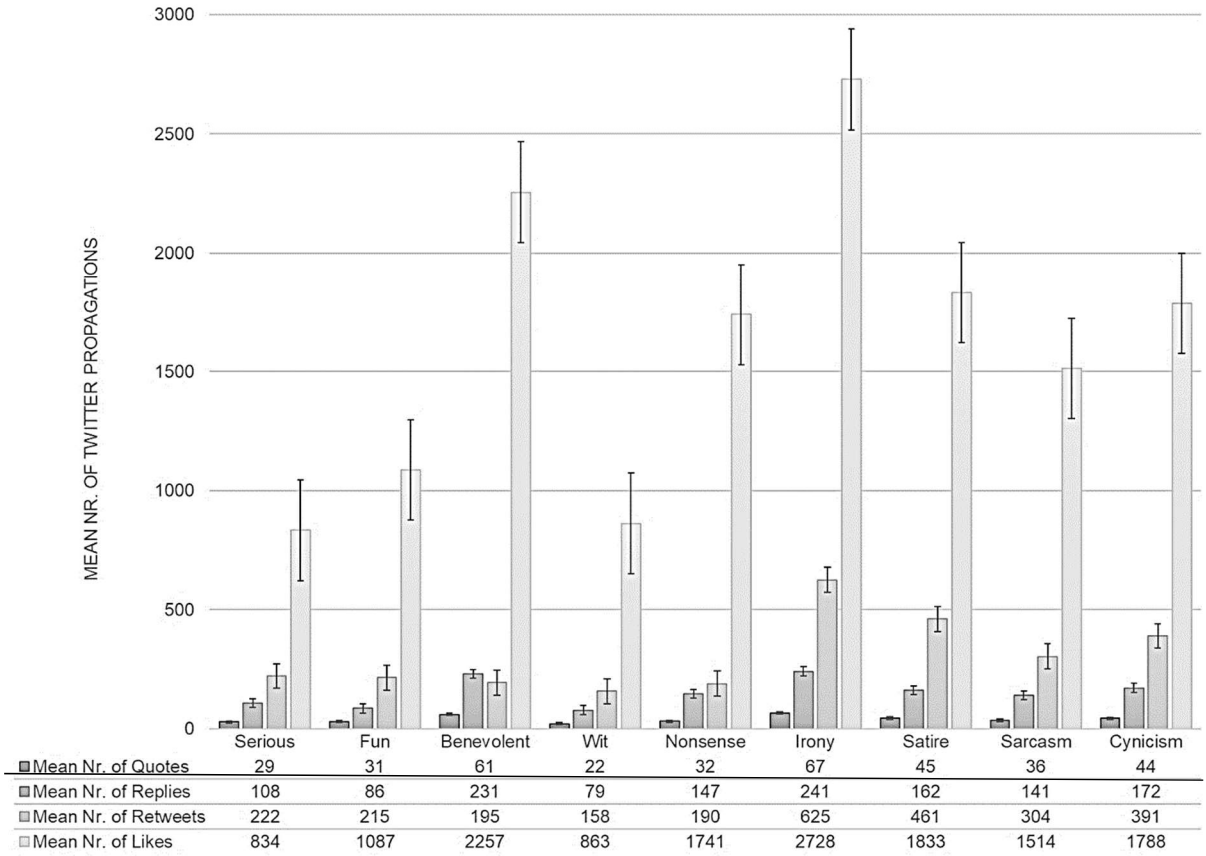
For both A1) and A2) statistically significant effects for humour-degree were determined, regarding the propagation $F (4, 25831) = 46.61$, $p < .001$, Eta-squared = .007. For A2) and regarding binary humour-degree in respect of propagation $F (1, 225005) = 1836.43$, $p < .001$, Eta-squared = .008.

A Bonferroni post-hoc test indicated statistically significant differences between each of the humour-degrees. The highest mean for the politicians was “hilarious” ($M = .61$, $SD = 2.30$, $n = 7802$) and the largest difference was found between “wry smile” and “hilarious” ($M_{difference} = .47$, $p <.001$). The assumption of equal variances in the population for the humour-degree for the politicians was again violated, Levene’s $F(4, 25831) = 101.68$, $p <.001$.

The pattern for degree of humour is visualized clearly in figure 8. Humour intensification stimulates virality and this is true for all types of propagation with just one exception (retweets of “smile”). The overall proportion of non-humorous tweets (figure 9) confirms well the result from the humour-types (88.52% versus 91.56%). For the binary analysis of degree of humour this is supported through the ANOVA. The highest mean for the politicians showed the highest impact for humorous ($M = .38$, $SD = 1.94$, $n = 25836$). Of note: the assumption of equal variances in the population was violated for the politicians, Levene's $F (1, 225005) = 3371.23$, $p <.001$.

Figure 6

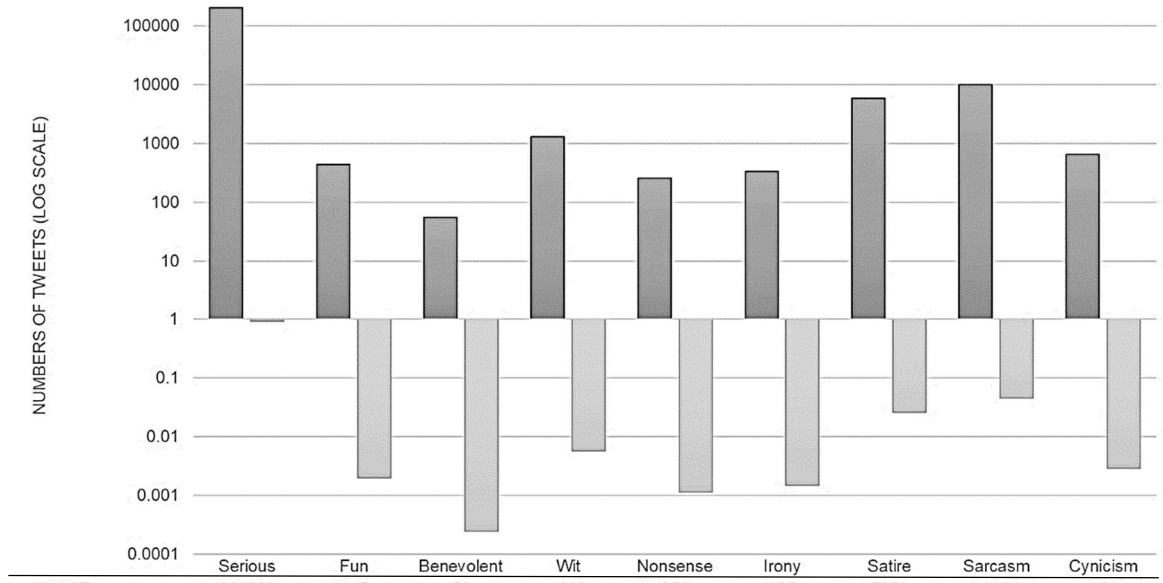
Politicians: Mean Propagations per Humour-type (SE of Mean)



	Serious	Fun	Benevolent	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Mean Nr. of Quotes	29	31	61	22	32	67	45	36	44
Mean Nr. of Replies	108	86	231	79	147	241	162	141	172
Mean Nr. of Retweets	222	215	195	158	190	625	461	304	391
Mean Nr. of Likes	834	1087	2257	863	1741	2728	1833	1514	1788

Figure 7

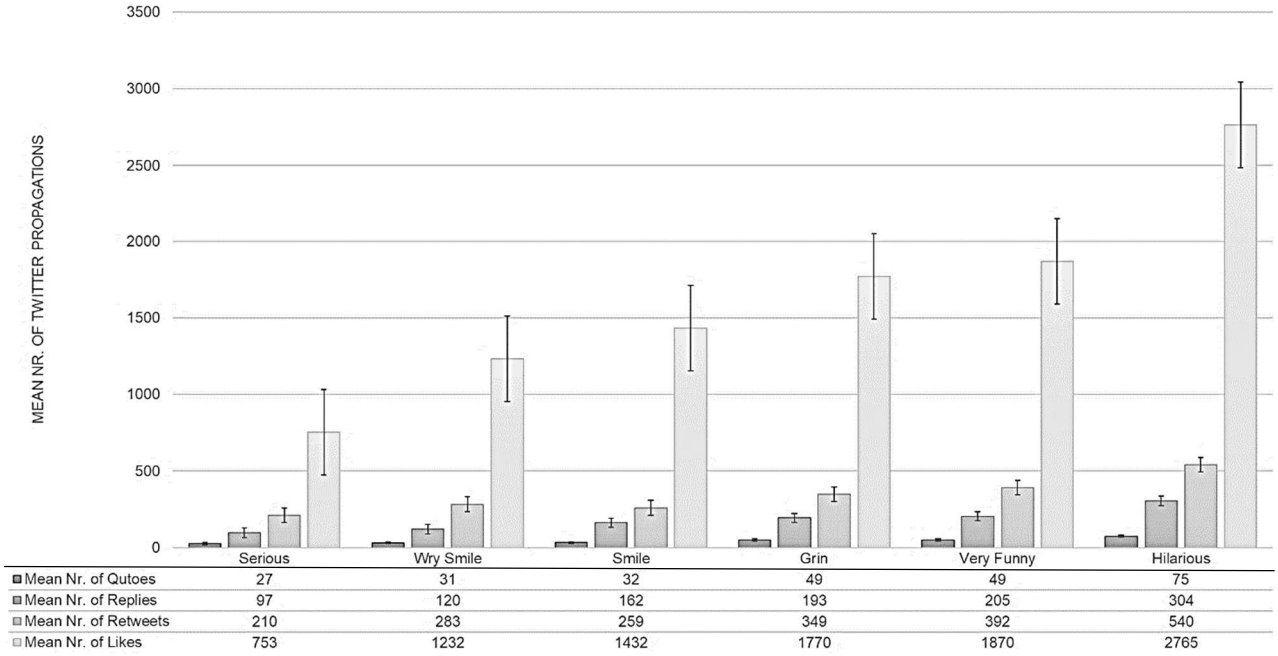
Politicians: Numbers of Tweets per Humour-type



	Serious	Fun	Benevolent	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Nr. of Tweets	206089	445	56	1292	259	337	5858	10104	651
Proportion of Tweets	91.56%	0.20%	0.02%	0.57%	0.12%	0.15%	2.60%	4.49%	0.29%

Figure 8

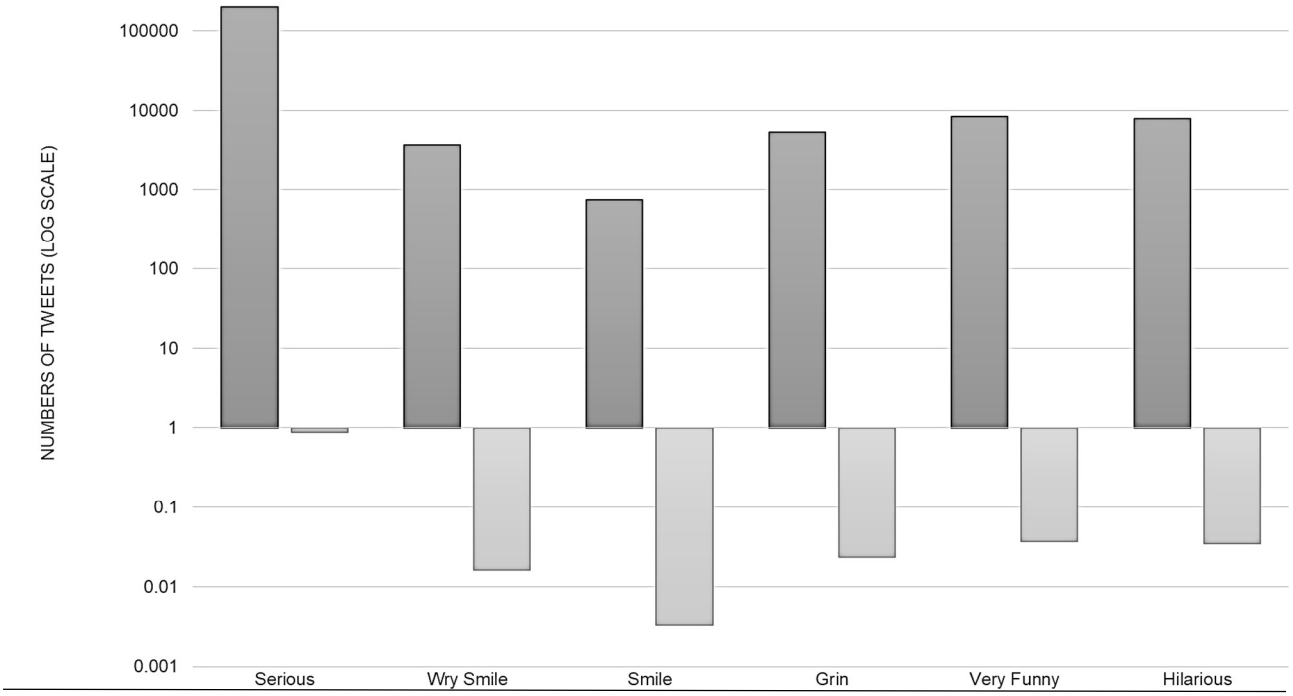
Politicians: Mean Propagations per Humour-Degree (SE of Mean)



	Serious	Wry Smile	Smile	Grin	Very Funny	Hilarious
Mean Nr. of Qutoes	27	31	32	49	49	75
Mean Nr. of Replies	97	120	162	193	205	304
Mean Nr. of Retweets	210	283	259	349	392	540
Mean Nr. of Likes	753	1232	1432	1770	1870	2765

Figure 9

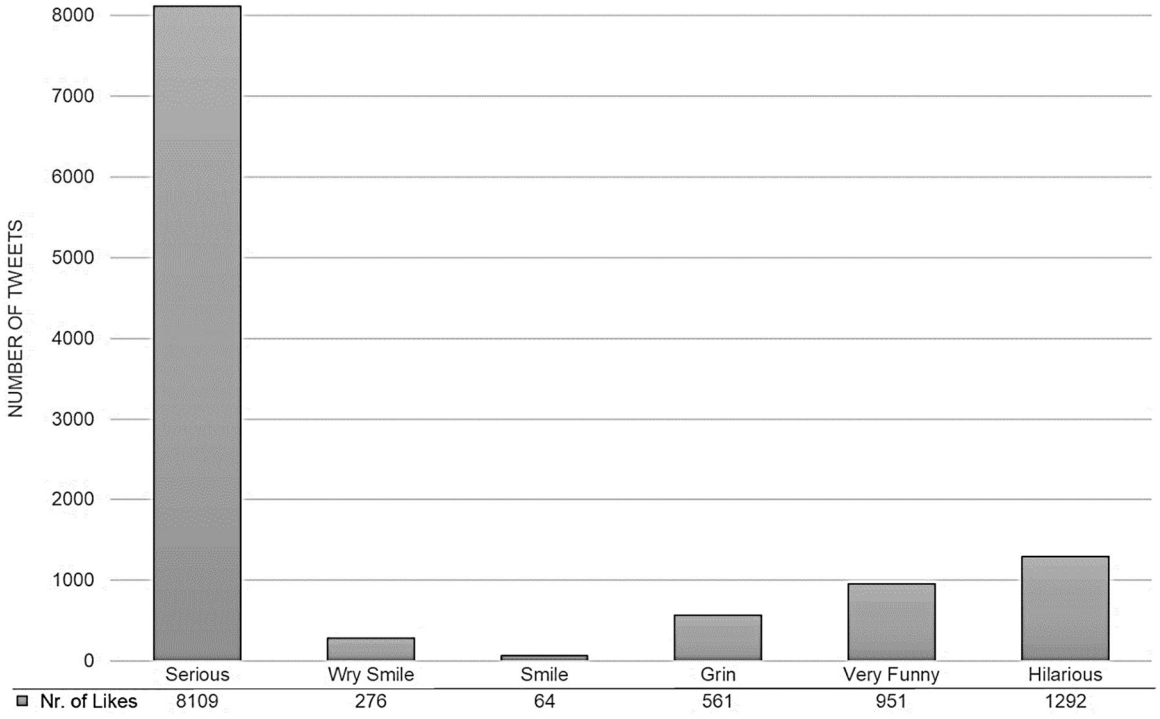
Politicians: Numbers of Tweets per Humour-Degree



	Serious	Wry Smile	Smile	Grin	Very Funny	Hilarious
Nr. of Tweets	199249	3666	745	5283	8345	7804
Proportion of Tweets	88.52%	1.63%	0.33%	2.35%	3.71%	3.47%

Figure 10

Politicians: Distribution of Upper 5% of Liked Tweets



Political Journalists

As with politicians, tweets (N=194,935) of 232 political journalists were collected for the same 2 year period. Referring to the graphical results in Appendix D, the picture for journalists (figure D1) is similar in some respects to politicians:

- The darker humour-types find more resonance except regarding quoting which is quite uniform across all humour-types, but with benevolence being proportionately most quoted.
- Generally speaking, propagation increases as humour is intensified (figure D3)
- In all cases but one (benevolence) there are more likes than retweets, more retweets than replies and more replies than quotes

Humour-Type. The Bonferroni post-hoc test indicated for cynicism the greatest impact on virality (M = .29, SD = 1.48, n = 387) of which the difference between serious and cynicism is the most pronounced ($M_{difference}$ = -.37, p <.001). A statistically significant effect of

the humour-type was determined regarding the propagation F (9, 194816) = 51.58, p < .001, Eta-squared = .002.

Humour-Degree. In step A1) a significant effect of the humour degree was found regarding the propagation F (4, 9559) = 16.17, p < .001, Eta-squared = .007. The Bonferroni post-hoc test for the journalists supports the finding, that there are several statistically significant differences found between each of the humour-degrees. The highest mean for the journalists was “hilarious” (M = .19, SD = 1.20, n = 1894), while the biggest difference in degree of humour was between “smile” and “hilarious” ($M_{difference}$ = .23, p <.001). The assumption of equal variances in the population for the humour-degree for the journalists, was also violated here, Levene’s F (4, 9559) = 33.15, p <.001.

In the second step A2) a statistically significant effect of the binary humour-degree was determined regarding the propagation F (1, 194929) = 488.24, p < .001, Eta-squared = .002. The results of the post-hoc test support the finding that humour as opposed to seriousness shows a higher level of propagation. The highest mean for the journalists was hilarious (M = .06, SD = .89, n = 9564). The assumption of equal variances in the population was however violated, Levene's F (1, 194929) = 855.20, p <.001.

Levene’s F. In all the analyses conducted for H2 and H3 it was noted that the Levene’s F was significant. The hypothesis that the variances in the groups are equal had therefore to be rejected. However, the tested samples all contain natural groupings based on the tweets of politicians and political journalists. Grace-Martin (2020) explains that the resulting power is based on the smaller sample size. Though it must be noted here that while the sample size may be relatively “small”, especially compared to the serious group, the smallest group still contains substantial numbers of tweets (cp. Figure 7, 9, D2 and D4). See limitations section for a further discussion.

Summary of Results

It would appear that, with some marginal exceptions, there is significantly more social media propagation when some kind of humour is attempted. The resonance for politicians is

especially positive for irony (all types of propagation) and in likes and replies to benevolent humour. For the darker humour-types, the propagation is roughly double that of serious tweets.

Where the journalists differ somewhat from the politicians is that their cynicism finds more resonance as opposed to politicians' irony. Benevolence from journalists is less appreciated than when it comes from politicians. The popularity by degree of humour for journalists displays a very similar profile to that of the politicians. More journalistic humour stimulates virality for all types of propagation with, once again, just the exception of "smile" (retweets and likes).

One notices (figure D2) that political journalists tweet with even less humour overall than politicians (< 4%) but, as with politicians (figure 8), use rather more satire and sarcasm. Even though it is greeted with less resonance, journalists are more frequently benevolent in their humour than politicians. The degree of humour from journalists is also very similarly distributed as with politicians, but is overall more sparingly used (figure D3).

The overall proportion of non-humorous tweets (figure D4) confirms almost exactly the result from the humour-types (96.23% versus 95.09%). Important to remember here is, that the results stem from fine-tunings based on 2 completely different datasets and NLMs, yet agree remarkably closely. Also worth mentioning is that higher degrees of humour are applied a little more often than weaker degrees.

Comedians. As mentioned in the Method section, a third control group of comedic tweets were analysed in the same way as for politicians and journalists. The complete results are available for comparison in Appendix E, but briefly stated, comedians are (much) more often (much) funnier, than both politicians and journalists. They are largely sarcastic with much wit and nonsense, but they also like to use satire. Here again, consumers seem to react most to the darker humour-types, but once again strong resonance for benevolent humour is observed.

Conclusion and Discussion

This study aimed to implement an optimized humour detection mechanism based on neural language models (NLMs) using the denoising autoencoder (AE) and autoregressive (AR) patterns, thereby solving 2 tasks discerning between a) 9 humour-types and b) 6 degrees of humour. Using automated humour-categorization, it should then be possible to discern the effect of humour on the propagation of political messaging in social media networks for different categories of actor.

The results of the project let us conclude that NLMs – with appropriate fine-tuning – do enable good automated differentiation between the 9 chosen humour-types and a moderately good estimation of the degree of humour in a text. It was additionally possible to replicate results of previous researchers of the simpler binary humour-categorization task (see Appendix C), the results obtained being in agreement (F1 and MCC > .99) with similar approaches by Annamoradnejad & Zoghi (2021) and Weller and Seppi (2019). RQ1 can be answered in the affirmative.

It has become clear that the GPT-2-derived models (of type AR) are considerably less precise in predicting humour-type, although XLNet (also of type AR) produced thoroughly usable results (MCC=0.68, F1=0.81) whilst not amongst the very best at this task. In the humour-degree task the picture is similar, but at a generally lower level of Precision and Recall for all models. RQ2 cannot be answered completely in the affirmative, since the results of XLNet-derivates (see Table B1) are hardly significantly worse than the AE models. It can merely be concluded that the GPT2-derivates are less appropriate for both of the tasks at hand (concurring with expectations from the SuperGLUE benchmark[29]). H1 has, therefore, to be rejected.

[29] SuperGLUE (https://super.gluebenchmark.com/leaderboard): benchmark test for evaluating general-purpose language understanding systems

Previous research examining comedic content has focused primarily on the medium of TV (Ödmark, 2021) but regarding online political messaging there are many interesting conclusions, which can now be drawn. As hypothesized – and similarly concluded by Georgalidou (2011) regarding prevalence of aggressive humour by politicians – the highest levels of social network virality for the politicians (all types of propagation) do stem from ironic humour (see the results of the ANOVA and Bonferroni post-hoc test). Although (other than serious tweets) politicians turn to sarcasm most often (figure 7) and employ benevolence and irony least. This is in itself “ironic” since these types would evoke most resonance.

Conversely, when regarding political journalists, the most viral type of humour by far is cynicism (see Bonferroni post-hoc). Interestingly, the favourite form of humour for the journalists by proportion of tweets is – as with politicians – sarcasm. If both political journalists and politicians are considered to be messaging politically, then there is no single most effective humour-type for both groups. H2 must therefore be rejected.

Looking at the results for degree of humour it should be remembered that the confidence in the predictions are much weaker than in the other task ($F1 < .57$, $MCC < .43$). Notwithstanding this, there is an almost remarkably clear picture indicating that more intense humour enhances social media propagation. This trend is also observed perfectly by the control group of comedians (figure D3), although – very predictably – this group produces much higher proportions of humour generally and especially humour of higher degrees (figure D4). RQ3 can, in respect to level of humour, be affirmed (in Bonferroni post-hoc tests, the highest level of humour was always best), but regarding type of humour, the effectiveness varies from group to group. RQ3 has thus, overall, to be rejected.

Finally, there is the hypothesis (H3) that more intense levels of humour improve mean propagation – which is true for all groups (see figures 8, D3 and E3 as well as the results of the ANOVA, step A1) – but also that serious messaging, albeit in smaller proportions, is frequently quite effective (see ANOVA, step A2). This was made transparent by limiting the analysis to only the 5% most liked tweets for each group. Figures 10, D5 and E5

demonstrate that serious messaging in fact constitutes a majority of the most propagated tweets. This is even true for comedians, although not as markedly as for politicians or especially journalists. The reticence of (UK) journalists to use humour may relate to “the centrality of impartiality in the UK” (Bailey, 2018; Section 4: Impartiality - Introduction, n.d.). H3 is supported and it is clear that – as expected – non-humorous political messaging also contributes to virality.

Discussion

The project can clearly be split into two halves: In phase 1 the fine-tuning of multiple NLMs posed many problems of (training) data quality, correct parametrisation of the machine learning and performance considerations of the language models. Appropriate quality assessment techniques were required and some excellent tools could be employed at very low cost. In phase 2, applying the humour-detection software to large numbers of political messages and interpreting the results was, by comparison, uncomplicated and shows how, with a well-performing, pre-trained and fine-tuned NLM, many previously almost unassailable tasks in communication science can be successfully tackled, as previously concluded by McClelland et al. (2020). The possibilities of applying essentially the same techniques to other linguistically complex problems (other than humour) are most encouraging.

Consequences for the Use of Humour in Politics

The results of the study are in many respects unsurprising, but the clarity of the statistics should be of great help for all political practitioners when considering their communications strategy. It is noteworthy that less than 9% of MP’s tweets are in any way amusing (figure 7). Taken in combination with H3, an increase in usage of humour would seem to be possible and no bad thing. The least number of politicians’ tweets were benevolent – a tiny proportion of 0.02%. Considering that benevolently funny tweets enjoy a proportionally very high virality (figure 6), this might be something that politicians could leverage. This agrees with the interpretation of self-deprecatory humour by Tesnohlidkova (2021). Self-deprecatory humour – a part of benevolent humour – aids in likeability and is positively correlated with persuasiveness (Hoption, Barling & Turner, 2013). However, there

is a possibility of it – with its negative undertones – backfiring and may reinforce perceptions of vulnerability, which is why, according to McAndrew (2019), this type of humour is seldom used. Benevolent humour is particularly relevant for politicians as actors in the public eye to create a favourable image. Showing understanding for the problems and shortcomings of others in a sympathetic, humorous way can promote connections with the general public (Deen, 2018). Deen (2018) also argues that benevolent humour aides in cultivating a consciousness of the limitations of politics.

For political journalists the learnings point in a slightly different direction, namely towards darker and particularly more cynical humour – at least for those wishing to raise their profile in the social media (see also Brants et al. (2010) with regard to pure, non-humorous cynicism of journalists). It would seem, though, that they should leave benevolence to the politicians. But since journalists are often reacting to political activity, media outlets may be more intrigued by the possibility of quickly and automatically recognising humorous or incongruous framing emanating from leading politicians (cp. Cappella & Jamieson, 1996) – a facility easily achievable as a by-product of this project.

The tools developed here are believed to be the first of their kind to allow quantification of humour at scale and as such could be included in communications processes on an ongoing basis. Knowing that “In the United States, political jokes on television are monitored by politicians to gauge the success or failure of their campaigns” (Lockyer & Pickering, 2009) this new possibility of automating humour measurement has a natural first application.

Limitations and Future Research

Despite moderate to good correlations of results to test data – particularly in regard of the humour-type categorisation task (MCC > 0.7) – the precision of the fine-tuned models could still be usefully improved. It was deduced that the quality of the results (under the pre-condition that fundamentally appropriate hyperparameters for learning are used) is more attributable to the quality of the data-annotation than to huge amounts of training data.

Improving Training Data. In combination with the same data-procurement techniques employed here, it would be unproblematic using the now available HDS to pre-sort large quantities of data, which would markedly simplify the task of an annotator, who could then quite quickly work through a batch of, for example, satirical texts and sort them into merely satirical or not. This would be a mentally and linguistically much less demanding procedure than categorising one text after another into one of 9 different types.

Levene's F. Since there are only 2 groups from 30 that even have a sample size less than 100 – benevolence for politicians and wit for political journalists – the slight issue with Levene's F significance could potentially be mitigated (as mentioned in the method section) by collapsing these categories into the fun category. A preferred approach would, however, involve improvement of training data (see above) and an even higher powered data procurement.

NLM Performance and Limitations. The confusion matrix in figure G2 demonstrates the limitations of small/tiny models, where Ernie-tiny (trained with 1 or 3 epochs) could not correctly detect ***any*** satire, benevolence, fun or cynicism. It would seem that fewer layers would limit the number of learnable categories for the model, Ernie-tiny having only 3 hidden layers, whereas the other models have 6, 12 or in the case of XLNet-Large and Ernie-Large 24.

Ensembling. Combining multiple models (of which this project has already generated many) to get a "best-of" result on inference is a well-documented technique in the literature for improving precision. Xu, Barth and Solis (2019) exploit the relative strengths of multiple BERT models in ensembles, whilst Huang, She and Zhang (2020) employ multi-class boosting to BERT models. Bagging, boosting and stacking are three of the most widely used ensemble types, which might be implemented.

Model Size. Even as CPU (or GPU) and RAM capacities grow, NLMs are voracious consumers of resources. The experience of this project is not necessarily that very large models always have to be used, but they do display some enhanced possibilities and it is thought that with more categories and for multi-task training (which was not attempted, here) the results will be better with more heads and more hidden layers. The recommendation

would be that in the development phase of a project, whilst training and test data is being assembled, it makes sense to use a faster, lightweight model such as DistilBERT. To optimize the quality (Precision and Recall) of the fine-tuning, the completed training data can then be applied to one of the larger models. There are however risks with training ever larger models, such as environmental or financial costs, as well as with too uniform data introducing an encoding bias, which in later use might cause harm (Bender et al., 2021).

If inference performance and/or model size is an issue for applications, there are optimization techniques, which can be applied. Pruning is a method, long in existence with neural networks (Wang, Wohlwend & Lei, 2020) where it has been demonstrated that accepting a performance drop of between 1% and 3% a reduction in model size of 75% (for BERT and XLNet) or 33% for DistilBERT is achievable (Sajjad et al. 2020). Alternatively, if massively large models are used (e.g. GPT 3 or Switch-C) there are techniques to extract the task-specific knowledge from the massive model by transfer learning to a small model such as DistilBERT (Sanh et al., 2020). The resulting lightweight model will not only be immensely smaller, but also possess much better performance. The down-side is that the smaller model will be very specialised, but this should not preclude use in specialised cases, such as in mobile-apps.

Humour-Degree vs. Type. The test results for measurement of humour-degree are considerably poorer than those for humour-type. But why? On the one hand, the annotations for humour-degree were derived from Reddit online-voting and as such, certainly have legitimacy, but on the other hand, precisely deciding how humorous a text is, is a rather different task to deciding between two or more types of humour. Type of humour has more to do with certain language patterns than does the degree of humour, which ultimately is very reliant on the cleverness of the message.

My interpretation of the research is that NLMs trained for these tasks are really quite good at recognising humorous language patterns, but interpreting cleverness is still a huge challenge. The difference is analogous to a political speech-writer, who on the one hand should include the important policy points being “sold” by the politician, but ideally also

packages these points using rhetorical devices (Bull & Miskinis, 2015). It seems probable that the methods presented here successfully recognise the stylistic patterns of humourists without necessarily deciphering the inner meaning in every respect. Nonetheless, this ability to recognise reliably “the packaging” is a definite and important step forward in automated content analysis.

Notes

1. Much useful data was created in this project for which there was no space or time to develop further. For every inference from the fine-tuned models the confidence score of the classification was produced and saved (these scores are available in https://github.com/jb-diplom/humour-detection/results/correlations). It would provide insight to drill down to tweets of particular humour-types with the very highest scores.
2. Noticeably, in the present analysis the motivation behind the humorous tweets was not considered. In a next step a signed network analysis could be conducted, to further analyse the implications of the humour.
3. The user identities are also available in all cases, so an analysis on a person-by-person basis would be equally possible. It would be of interest to know how densely, particular types of humour are concentrated around individual politicians and journalists.
4. In retrospect it would be good (and a very simple task) to repeat the analyses of journalists, but to omit all tweets from organisations, since it is thought that most organisations deliberately avoid whimsical copy.
5. In Appendix G some learnings and observations from the project regarding hyperparameter tuning are summarized.

Funding

The research was supported by a funding grant from the Digital Communication Methods Lab (https://www.digicomlab.eu).

Bibliography

Annamoradnejad, I., & Zoghi, G. (2021). ColBERT: Using BERT Sentence Embedding for Humor Detection. *ArXiv:2004.12765 [Cs]*. http://arxiv.org/abs/2004.12765

Badlani, R., Asnani, N., & Rai, M. (2019). An Ensemble of Humour, Sarcasm, and Hate Speechfor Sentiment Classification in Online Reviews. *Proceedings of the 5th Workshop on Noisy User-Generated Text (W-NUT 2019)*, 337–345. https://doi.org/10.18653/v1/D19-5544

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*. ArXiv:1409.0473 [Cs, Stat]. http://arxiv.org/abs/1409.0473

Bailey, R. (2018). When journalism and satire merge: The implications for impartiality, engagement and 'post-truth' politics – A UK perspective on the serious side of US TV comedy. *European Journal of Communication*, *33*(2), 200–213. https://doi.org/10.1177/0267323118760322

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? . *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610–623. https://doi.org/10.1145/3442188.3445922

BhMadStudio. (2017, February 6). *#HashtagWars: Learning a Sense of Humor*. SemEval-2017 Task 6. https://alt.qcri.org/semeval2017/task6/index.php?id=data-and-tools

Brants, K., de Vreese, C., Möller, J., & van Praag, P. (2010). The Real Spiral of Cynicism? Symbiosis and Mistrust between Politicians and Journalists. *The International Journal of Press/Politics*, *15*(1), 25–40. https://doi.org/10.1177/1940161209351005

Brown, J. (1995, September / October). Funny you should say that: Use humor to help your students. *Creative Classroom*, 10, 80-81

Brownlee, J. (2018, May 23). A Gentle Introduction to k-fold Cross-Validation. *Machine Learning Mastery*. https://machinelearningmastery.com/k-fold-cross-validation/

Bull, P., & Miskinis, K. (2015). Whipping It Up! An Analysis of Audience Responses to Political Rhetoric in Speeches From the 2012 American Presidential Elections. *Journal of Language and Social Psychology*, *34*(5), 521–538. https://doi.org/10.1177/0261927X14564466

Buzzberry, C. (2020, December 14). Huge list of British Comedians on Twitter to follow for laughs! [Blog]. Chucklebuzz. https://chucklebuzz.com/blog/british-comedians-on-twitter-follow-laughs/83428/

Boukes, M., Boomgaarden, H. G., Moorman, M., & de Vreese, C. H. (2015). At Odds: Laughing and Thinking? The Appreciation, Processing, and Persuasiveness of Political Satire: At Odds: Laughing and Thinking? *Journal of Communication*, *65*(5), 721–744. https://doi.org/10.1111/jcom.12173

Cappella, J. N., & Jamieson, K. H. (1996). News Frames, Political Cynicism, and Media Cynicism. *The Annals of the American Academy of Political and Social Science*, *546*, 71–84. JSTOR.

Carletta, J. (1996). Assessing agreement on classification tasks: The kappa statistic. *ArXiv:Cmp-Lg/9602004*. http://arxiv.org/abs/cmp-lg/9602004

Cheng, D., Chan, X. W., Amarnani, R. K., & Farivar, F. (2021). Finding humor in work–life conflict: Distinguishing the effects of individual and co-worker humor. *Journal of Vocational Behavior*, *125*, 103538. https://doi.org/10.1016/j.jvb.2021.103538

Davis, J. L., Love, T. P., & Killen, G. (2018). Seriously funny: The political work of humor on social media. *New Media & Society*, *20*(10), 3898–3916. https://doi.org/10.1177/1461444818762602

Milner Davis, J. (2003). *Farce*, 2nd Edn. Piscataway, NJ: Transaction Publishing.

Deen, P. (2018). Senses of Humor as Political Virtues: SENSES OF HUMOR AS POLITICAL VIRTUES. *Metaphilosophy*, *49*(3), 371–387. https://doi.org/10.1111/meta.12297

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. ArXiv:1810.04805 [Cs]. http://arxiv.org/abs/1810.04805

Fan, X., Lin, H., Yang, L., Diao, Y., Shen, C., Chu, Y., & Zou, Y. (2020). Humor detection via an internal and external neural network. *Neurocomputing*, *394*, 105–111. https://doi.org/10.1016/j.neucom.2020.02.030

Freud, S. (1928). *Humour*. The International Journal of Psychoanalysis, 9, 1–6.

Freeman, L. C. (1978). Centrality in social networks conceptual clarification. *Social Networks, 1*(3), 215–239. https://doi.org/10.1016/0378-8733(78)90021-7

Forabosco, G. (1992). Cognitive aspects of the humor process: The concept of incongruity. *Humor - International Journal of Humor Research, 5*(1–2). https://doi.org/10.1515/humr.1992.5.1-2.45

Hendriks, H., & Strick, M. (2019). A Laughing Matter? How Humor in Alcohol Ads Influences Interpersonal Communication and Persuasion. *Health Communication*, 1–9. https://doi.org/10.1080/10410236.2019.1663587

Georgalidou, M. (2011). Chapter 4. “Stop caressing the ears of the hooded”: Political humour in times of conflict. In V. Tsakona & D. E. Popa (Eds.), *Discourse Approaches to Politics, Society and Culture* (Vol. 46, pp. 83–107). John Benjamins Publishing Company. https://doi.org/10.1075/dapsac.46.07geo

Grace-Martin, K. (2020, December 18). *When Unequal Sample Sizes Are and Are NOT a Problem in ANOVA*. The Analysis Factor. https://www.theanalysisfactor.com/when-unequal-sample-sizes-are-and-are-not-a-problem-in-anova/comment-page-1/

Hennefeld, M. (2016, December 18). Laughter in the Age of Trump [A Critical Forum on Media and Culture]. *Flow.Journal*. http://www.flowjournal.org/2016/12/laughter-in-the-age-of-trump/

Heintz, S., Ruch, W., Aykan, S., Brdar, I., Brzozowska, D., Carretero-Dios, H., Chen, H.-C., Chłopicki, W., Choi, I., Dionigi, A., Ďurka, R., Ford, T. E., Güsewell, A., Isler, R. B., Ivanova, A., Laineste, L., Lajčiaková, P., Lau, C., Lee, M., … Wong, P. S. O. (2020). Benevolent and Corrective Humor, Life Satisfaction, and Broad Humor Dimensions: Extending the Nomological Network of the BenCor Across 25 Countries. *Journal of Happiness Studies*, *21*(7), 2473–2492. https://doi.org/10.1007/s10902-019-00185-9

Higgie, R. (2014). Kynical dogs and cynical masters: Contemporary satire, politics and truth-telling. *HUMOR*, *27*(2). https://doi.org/10.1515/humor-2014-0016

Hoption, C., Barling, J., & Turner, N. (2013). "It's not you, it's me": Transformational leadership and self-deprecating humor. *Leadership & Organization Development Journal*, *34*(1), 4–19. https://doi.org/10.1108/01437731311289947

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 328–339. https://doi.org/10.18653/v1/P18-1031

Huang, T., She, Q., & Zhang, J. (2020). BoostingBERT:Integrating Multi-Class Boosting into BERT for NLP Tasks. *ArXiv:2009.05959 [Cs]*. http://arxiv.org/abs/2009.05959

Jenkins, H., Ford, S., & Green, J. (2013). *Spreadable media: Creating value and meaning in a networked culture*. New York University Press.

Joshi, R. (2016, September 9). Accuracy, Precision, Recall & F1 Score: Interpretation of Performance Measures. *Exsilio Solutions*. https://blog.exsilio.com/all/accuracy-precision-recall-f1-score-interpretation-of-performance-measures/

Joshi, A., Bhattacharyya, P., & Carman, M. J. (2016). Automatic Sarcasm Detection: A Survey. ArXiv:1602.03426 [Cs]. http://arxiv.org/abs/1602.03426

Johnson, A. J., & Mistry, K. (2013). The effect of joke-origin-induced expectancy on cognitive humor. *Humor*, *26*(2). https://doi.org/10.1515/humor-2013-0011

JustAnotherArchivist. (2020). Snscrape (0.3.4.) [Python 3.6]. https://github.com/JustAnotherArchivist/snscrape

Katz, E., Lazarsfeld, P. F. & Roper, E. (1955). Personal influence: The part played by people in the flow of mass communications. Glencoe (Ill.): Free press. http://lib.ugent.be/catalog/rug01:000039958

Katz, E. (1957). The Two-Step Flow of Communication: An Up-To-Date Report on an Hypothesis. *Public Opinion Quarterly*, *21*(1, Anniversary Issue Devoted to Twenty Years of Public Opinion Research), 61. https://doi.org/10.1086/266687

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (2nd ed). Sage.

Knoblock, N. (2016). Sarcasm and Irony as a Political Weapon: Social Networking in the Time of Crisis. In D. O. Orwenjo, O. Oketch, & A. H. Tunde (Eds.), *Political discourse in emergent, fragile, and failed democracies* (pp. 11–33). Information Science Reference.

Kopper, A. (2020). *The use of humour in diplomatic tweets: The affiliative potential of ridicule*. Cooperation and Conflict, 001083672097545. https://doi.org/10.1177/0010836720975458

Lauer, W. (1974). *Humor als Ethos: Eine moralpsychologische Untersuchung*. H. Huber.

Lazarsfeld, P. F., Berelson, B., & Gaudet, H. (1944). *The people's choice: How the voter makes up his mind in a presidential campaign* (Legacy edition). Columbia University Press.

Lihala, A. (2019, March 29). Attention and its Different Forms: An overview of generalised attention with its different types and uses. *Towards Data Science*. https://towardsdatascience.com/attention-and-its-different-forms-7fc3674d14dc

Lockyer, S., & Pickering, M. (Eds.). (2009). *Beyond a joke: The limits of humour*. Palgrave Macmillan.

Martin, R. A., & Lefcourt, H. M. (1983). Sense of humor as a moderator of the relation between stressors and moods. *Journal of Personality and Social Psychology*, *45*(6), 1313–1324. https://doi.org/10.1037/0022-3514.45.6.1313

Martin, R. A., Puhlik-Doris, P., Larsen, G., Gray, J., & Weir, K. (2003). Individual differences in uses of humor and their relation to psychological well-being: Development of the Humor Styles Questionnaire. *Journal of Research in Personality*, *37*(1), 48–75. https://doi.org/10.1016/S0092-6566(02)00534-2

McAndrew, F. T. (2019, September 21). *Politicians don't seem to laugh at themselves as much anymore*. The Conversation. https://theconversation.com/politicians-dont-seem-to-laugh-at-themselves-as-much-anymore-122103

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2020). Extending Machine Language Models toward Human-Level Language Understanding. *ArXiv:1912.05877 [Cs]*. http://arxiv.org/abs/1912.05877

McCloskey, M., & Cohen, N. J. (1989). Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation* (Vol. 24, pp. 109–165). Elsevier. https://doi.org/10.1016/S0079-7421(08)60536-8

Mendiburo-Seguel, A., & Heintz, S. (2020). Comic styles and their relation to the sense of humor, humor appreciation, acceptability of prejudice, humorous self-image and happiness. *HUMOR*, *33*(3), 381–403. https://doi.org/10.1515/humor-2018-0151

Moldovan, S., Muller, E., Richter, Y., & Yom-Tov, E. (2017). Opinion leadership in small groups. *International Journal of Research in Marketing*, *34*(2), 536–552. https://doi.org/10.1016/j.ijresmar.2016.11.004

Mukhongo, L. L. (2020). Participatory Media Cultures: Virality, Humour, and Online Political Contestations in Kenya. *Africa Spectrum*, *55*(2), 148–169. https://doi.org/10.1177/0002039720957014

Ödmark, S. (2021). Making news funny: Differences in news framing between journalists and comedians. *Journalism*, *22*(6), 1540–1557. https://doi.org/10.1177/1464884918820432

Plester, B. A., & Sayers, J. (2007). "Taking the piss": Functions of banter in the IT industry. *Humor – International Journal of Humor Research*, *20*(2). https://doi.org/10.1515/HUMOR.2007.008

Peterson, C., & Seligman, M. E. P. (2004). *Character strengths and virtues: A handbook and classification*. American Psychological Association ; Oxford University Press.

Potash, P., Romanov, A., & Rumshisky, A. (2017). SemEval-2017 Task 6: #HashtagWars: Learning a Sense of Humor. *Proceedings of the 11th International Workshop on Semantic Evaluation.(SemEval-2017)*, 49–57. https://doi.org/10.18653/v1/S17-2004

Ruch, W., Heintz, S., Platt, T., Wagner, L., & Proyer, R. T. (2018). Broadening Humor: Comic Styles Differentially Tap into Temperament, Character, and Ability. *Frontiers in Psychology*, 9, 6. https://doi.org/10.3389/fpsyg.2018.00006

Ruder, S. (2021, February 24). Recent Advances in Language Model Fine-tuning. Sebastian Ruder. https://ruder.io/recent-advances-lm-fine-tuning/

Riquelme, F., & González-Cantergiani, P. (2016). Measuring user influence on Twitter: A survey. *Information Processing & Management*, *52*(5), 949–975. https://doi.org/10.1016/j.ipm.2016.04.003

Sajjad, H., Dalvi, F., Durrani, N., & Nakov, P. (2021). On the Effect of Dropping Layers of Pre-trained Transformer Models. *ArXiv:2004.03844 [Cs]*. http://arxiv.org/abs/2004.03844

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter. *ArXiv:1910.01108 [Cs]*. http://arxiv.org/abs/1910.01108

Section 4: Impartiality—Introduction. (n.d.). BBC: Editorial Guidelines. Retrieved 10 June 2021, from https://www.bbc.co.uk/editorialguidelines/guidelines/impartiality

Tesnohlidkova, O. (2021). Humor and satire in politics: Introducing cultural sociology to the field. *Sociology Compass*, *15*(1). https://doi.org/10.1111/soc4.12842

The Hugging Face Team. (n.d.). Summary of the models. Huggingface. Retrieved 28 March 2021, from https://huggingface.co/transformers/model_summary.html

Vance, J. (2013). *An Evaluative Review of the Pragmatics of Verbal Irony* [MPhil thesis, University of Sheffield]. https://etheses.whiterose.ac.uk/3383/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. ArXiv:1706.03762 [Cs]. http://arxiv.org/abs/1706.03762

WaterCooler: Scraping an Entire Subreddit (/r/2007scape). (2019, March 18). *ORSBox*. https://www.osrsbox.com/blog/2019/03/18/watercooler-scraping-an-entire-subreddit-2007scape/

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2020). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. *ArXiv:1905.00537 [Cs]*. http://arxiv.org/abs/1905.00537

Wang, Z., Wohlwend, J., & Lei, T. (2020). Structured Pruning of Large Language Models. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6151–6162. https://doi.org/10.18653/v1/2020.emnlp-main.496

Weller, O., & Seppi, K. (2019). Humor Detection: A Transformer Gets the Last Laugh. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3619–3623. https://doi.org/10.18653/v1/D19-1372

Winter, S., & Neubaum, G. (2016). Examining Characteristics of Opinion Leaders in Social Media: A Motivational Approach. *Social Media + Society*, *2*(3), 205630511666585. https://doi.org/10.1177/2056305116665858

Xu, C., Barth, S., & Solis, Z. (2019). *Applying Ensembling Methods to BERT to Boost Model Performance*. Natural Language Processing with Deep Learning. https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1194/reports/default/15775971.pdf

Xu, W. W., Sang, Y., Blasiola, S., & Park, H. W. (2014). Predicting Opinion Leaders in Twitter Activism Networks: The Case of the Wisconsin Recall Election. *American Behavioral Scientist*, *58*(10), 1278–1293. https://doi.org/10.1177/0002764214527091

Appendix A

Humorous Text Discrimination Matrix

Table A1

Comic Styles Description Map

Description	Fun	Benevolent Humour	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Light Humour	X	X	X	X				
Dark humour					X	X	X	X
Superiority					X	X		
Critical					X	X	X	X
Ingroup vs Outgroup					X			
Evaluative Statement					X			
Target present					X			
Negative					X	X		X
Destructive								X
Disillusionment								X
Mockery						X	X	X
Moral values are ridiculous								X
Aggressive						X		X
Detecting Weakness						X		
Improving Humans						X		
Moral Criticism						X		
Hostility							X	
Ruthlessness							X	
Derisiveness							X	
Schadenfreude							X	
Ridicule of topics / people / institutions				X			X	
Accepting		X						
Sympathy		X						
Understanding		X						
Self-deprecation		X						
Aim is good mood	X							
Camaraderie	X							
Social	X							
Jovial	X							
Banter	X							
Unresolved Incongruities				X				
Embracing Absurdities				X				
Bizarre				X				
Fantastical				X				
Surprising			X					
Punchlines			X					
(Unusual) Combinations			X					

Appendix B

Results of Model Fine-Tuning

Table B1

Humour-Degree (k=6)

Model Type	Precision	Recall	F1	eval/accuracy	Train Epochs	Batch Size	Optimizer	fp16
electra-large-discriminator	0.621	0.522	0.56	0.515	5	16	Adafactor	false
xlnet-base-cased	0.57	0.552	0.555	0.548	5	12	AdamW	true
bert-base-uncased	0.558	0.55	0.552	0.538	5	32	AdamW	true
ernie-2.0-en	0.565	0.551	0.552	0.547	5	12	Adafactor	true
distilbert-base-uncased	0.565	0.551	0.552	0.545	10	12	Adafactor	true
xlnet-base-cased	0.593	0.542	0.551	0.542	3	32	AdamW	true
bert-base-uncased	0.562	0.545	0.55	0.544	3	32	Adafactor	true
distilroberta-base	0.566	0.538	0.546	0.541	5	128	Adafactor	true
distilbert-base-uncased	0.549	0.544	0.545	0.545	5	12	Adafactor	true
ernie-2.0-en	0.551	0.542	0.543	0.535	20	12	Adafactor	true
ernie-tiny	0.562	0.537	0.543	0.534	5	12	Adafactor	true
distilbert-base-uncased	0.545	0.535	0.538	0.548	3	12	AdamW	true
xlnet-base-cased	0.545	0.535	0.538	0.525	20	12	AdamW	true
distilbert-base-uncased	0.555	0.534	0.538	0.525	50	12	Adafactor	true
deberta-base	0.555	0.534	0.538	0.514	10	12	Adafactor	true
distilbert-base-uncased	0.538	0.536	0.536	0.534	20	12	Adafactor	true
deberta-base	0.554	0.531	0.536	0.535	10	12	Adafactor	true
ernie-2.0-large-en	0.555	0.529	0.535	0.543	5	8	AdamW	false
distilbert-base-uncased	0.534	0.531	0.531	0.529	10	12	Adafactor	true
distilbert-base-uncased	0.539	0.529	0.531	0.527	50	32	AdamW	true
nghuyong/ernie-2.0-en	0.536	0.527	0.531	0.5	5	12	Adafactor	true
distilbert-base-uncased	0.528	0.526	0.526	0.545	50	12	AdamW	true
distilbert-base-uncased	0.526	0.524	0.525	0.532	50	12	Adafactor	true
ernie-tiny	0.524	0.525	0.524	0.524	20	12	Adafactor	true
distilgpt2	0.51	0.329	0.319	0.53	10	16	Adafactor	true
distilgpt2	0.436	0.328	0.308	0.531	10	16	Adafactor	true
distilgpt2	0.563	0.197	0.218	0.528	5	16	AdamW	true
gpt2	0.25	0.237	0.203	0.539	20	16	Adafactor	false
gpt2	0.769	0.104	0.173	0.515	3	12	Adafactor	false
distilgpt2	0.342	0.15	0.169	0.529	20	16	Adafactor	true

Mean all (N=30)	0.539	0.474	0.479	0.533
SD (N=30)	0.082	0.131	0.127	0.011
Mean AE (n=21)	0.552	0.535	0.540	0.533
SD AE (n=21)	0.020	0.009	0.010	0.012
Mean GPT (n=9)	0.463	0.320	0.326	0.485
SD GPT (n=9)	0.141	0.165	0.157	0.009

Table B2

Humour-type (k=9)

Model Type	Precision	Recall	F1	eval/accuracy	Train Epochs	Batch Size	Optimizer	fp16
ernie-2.0-large-en	0.810	0.800	0.804	0.811	3	8	AdamW	true
ernie-2.0-large-en	0.807	0.802	0.803	0.816	3	16	AdamW	true
roberta-base	0.809	0.798	0.802	0.812	5	32	Adafactor	true
ernie-2.0-large-en	0.804	0.798	0.800	0.816	3	8	AdamW	true
distilbert-base-uncased	0.817	0.787	0.799	0.806	3	32	Adafactor	false
ernie-2.0-en	0.803	0.797	0.798	0.814	2	16	AdamW	true
distilbert-base-uncased	0.814	0.785	0.797	0.805	3	32	Adafactor	true
distilroberta-base	0.809	0.788	0.797	0.817	5	64	Adafactor	true
ernie-2.0-en	0.801	0.794	0.797	0.813	5	12	Adafactor	true
ernie-2.0-large-en	0.801	0.794	0.797	0.808	5	8	AdamW	false
ernie-2.0-en	0.813	0.788	0.796	0.808	1	16	AdamW	true
xlnet-base-cased	0.802	0.790	0.794	0.815	5	32	AdamW	true
distilroberta-base	0.801	0.786	0.792	0.814	10	64	Adafactor	true
xlnet-large-cased	0.799	0.787	0.792	0.806	5	12	AdamW	true
bert-base-uncased	0.797	0.787	0.791	0.815	5	32	Adafactor	true
bert-base-uncased	0.797	0.787	0.791	0.803	10	32	AdamW	true
distilroberta-base	0.794	0.786	0.789	0.800	20	128	Adafactor	true
electra-large-discriminator	0.790	0.789	0.789	0.806	5	16	Adafactor	true
ernie-2.0-en	0.789	0.789	0.788	0.800	10	12	Adafactor	true
distilbert-base-uncased	0.790	0.782	0.785	0.802	10	12	Adafactor	true
distilbert-base-uncased	0.790	0.782	0.785	0.802	10	12	Adafactor	true
electra-base-discriminator	0.782	0.782	0.781	0.798	20	12	Adafactor	true
ernie-tiny	0.843	0.736	0.777	0.748	5	128	AdamW	true
ernie-tiny	0.841	0.737	0.777	0.747	3	16	AdamW	true
ernie-tiny	0.847	0.733	0.775	0.743	3	128	AdamW	true
distilgpt2	0.996	0.537	0.697	0.806	10	16	Adafactor	false
distilgpt2	0.936	0.524	0.669	0.801	20	16	AdamW	true
distilgpt2	0.877	0.536	0.664	0.788	50	16	Adafactor	true
distilgpt2	0.928	0.513	0.657	0.804	20	16	AdamW	true
gpt2	0.728	0.562	0.618	0.790	20	12	AdamW	true

Mean all (N=30)	0.820	0.742	0.770	0.800
SD (N=30)	0.051	0.095	0.050	0.020
Mean AE (n=23)	0.806	0.783	0.792	0.800
SD AE (n=23)	0.017	0.019	0.008	0.022
Mean AR (n=7)	0.867	0.607	0.699	0.801
SD AR (n=7)	0.087	0.116	0.063	0.009

Appendix C

Binary Humour Detection

Binary Humour-Degree

Due to the very poor values achieved with k=6, it was decided to experiment with collapsing all degrees of humour into a single humour category to discover if at least a binary test could work well. An additional restructuring of the data was performed for the humour-degree (previously k=6) to fine-tune for a dichotomous classification (humorous/ non-humorous, k=2), enabling a comparison with other research such as that mentioned in the Introduction that had focused on this binary decision-making (e.g. Annamoradnejad and Zoghi in 2021).

The results were very good (cp. Table C1). All models including those with the GPT architecture produced F1 values >= 0.933. The non-GPT models producing identically good results for Precision, Recall and then of course for F1, too (F1=0.997). It can be concluded that the binary humour task is quite easily solved with the chosen methodology. The results are in agreement with similar approaches by Annamoradnejad and Zoghi (2021) and Weller and Seppi (2019).

Table C1

Precision metrics for all model architectures for, Humour-Degree (k=2)

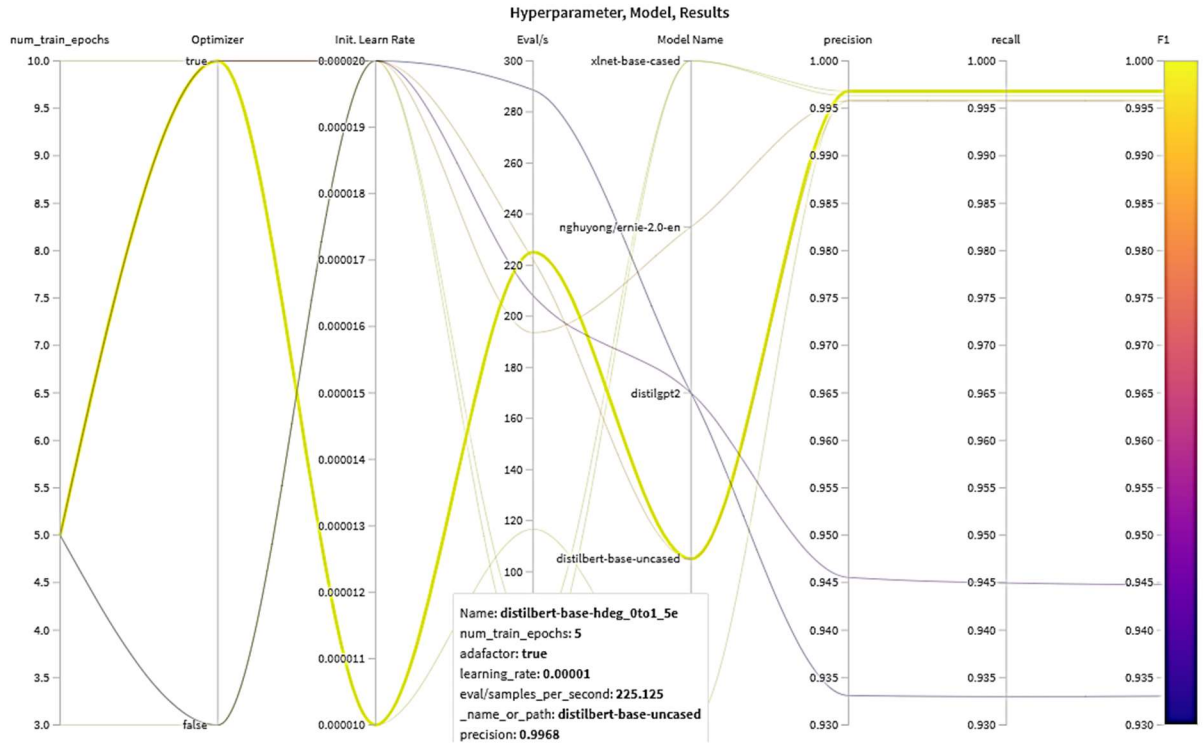
Model Type	Precision	Recall	F1	eval/accuracy	Train Epochs	Batch Size	Optimizer	fp16
xlnet-base-cased	0.997	0.997	0.997	0.997	5	12	AdamW	true
distilbert-base-uncased	0.997	0.997	0.997	0.997	5	12	Adafactor	true
bert-base-uncased	0.997	0.997	0.997	0.998	5	12	Adafactor	true
xlnet-base-cased	0.996	0.996	0.996	0.999	3	32	AdamW	true
distilbert-base-uncased	0.996	0.996	0.996	0.997	10	12	Adafactor	true
nghuyong/ernie-2.0-en	0.996	0.996	0.996	0.997	5	12	Adafactor	true
distilgpt2	0.946	0.945	0.945	0.995	5	32	Adafactor	false
distilgpt2	0.933	0.933	0.933	0.996	5	32	AdamW	false
Mean (N=8)	0.982	0.982	0.982	0.997				
SD	0.025	0.025	0.025	0.001				

Due to the uniformly high precision, testing was restricted to N=8 cases. For purely practical reasons (performance), it was decided to continue into the next phase of correlating

political tweets with the distilbert-base model (trained for 5 epochs). From the group of the very best models, DistilBERT demonstrated the best performance (225 evaluations/s) in for evaluation. The visualization in figure C1, highlighting the distilbert-base model, shows the equality of Precision and Recall in all cases. A distilbert-base model was additionally trained for 10 epochs and a higher initial learn rate, but achieved essentially identical results as with 5 epochs of training.

Figure C1

Hyperparameter Choice for Binary Humour-Degree

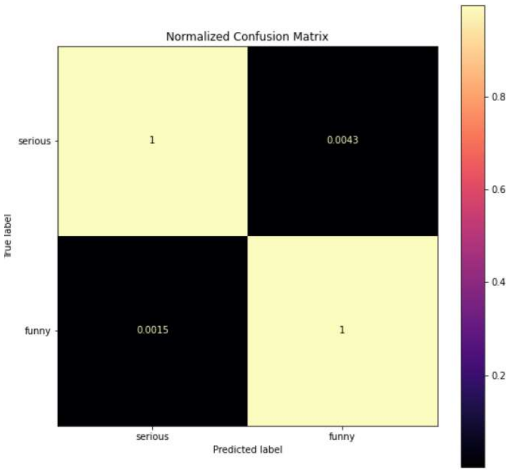


Note: Interactive plot available at: [Project Dashboard (wandb.ai)]

The normalized confusion matrix for the distilbert-base is shown in figure C2. The matrix is unremarkable, given the F1 value of 0.997 - almost no FPs or FNs, based on the used test-data (N=4,000).

Figure C2

Normalized Confusion Matrix for Distilbert-base (k=2)



As with the confusion matrix, the ROC and Precision/Recall graphs are, as expected almost perfect.

Figure C3

ROC and Precision/Recall Graphs for Distilbert-base (k=2)

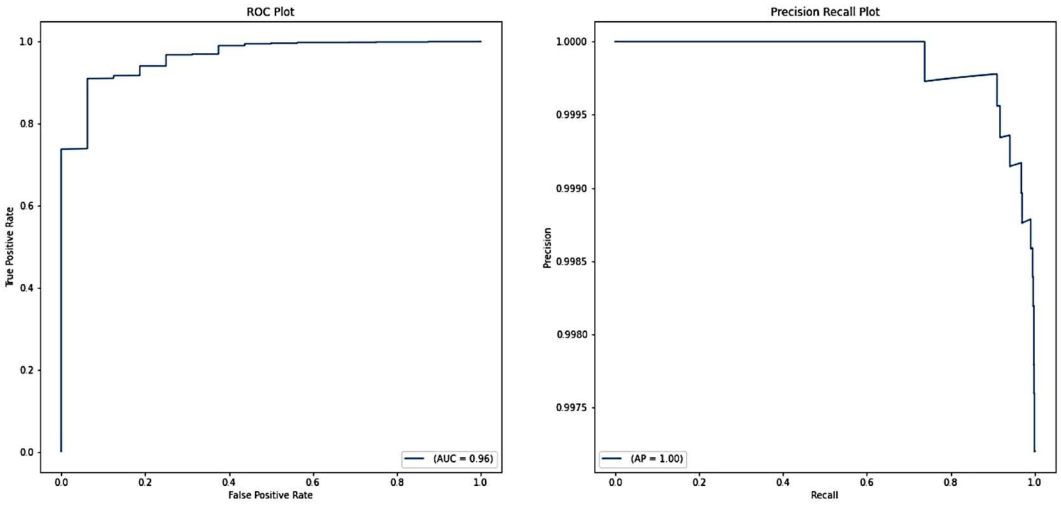
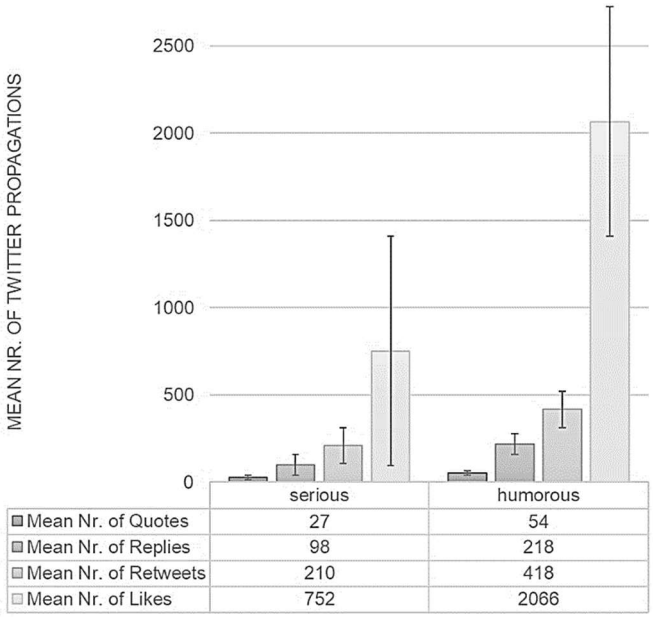


Figure C4

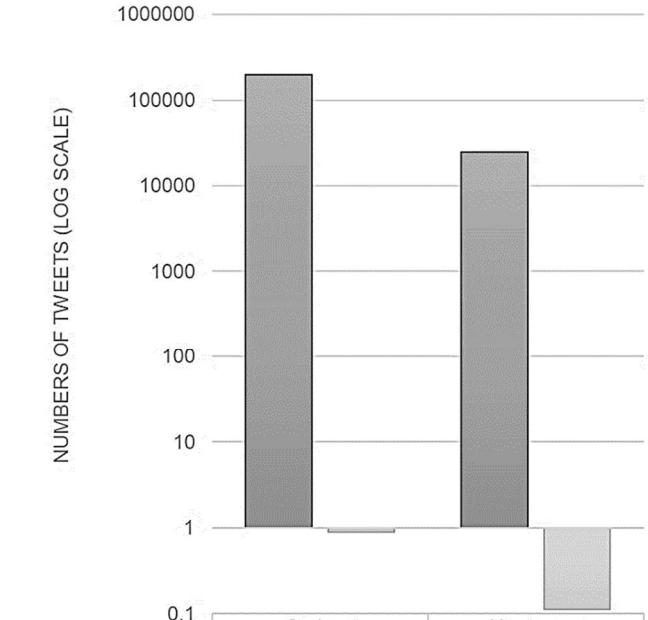
Politicians: Mean Propagations per Binary Humour-Degree (SE of Mean)



	serious	humorous
Mean Nr. of Quotes	27	54
Mean Nr. of Replies	98	218
Mean Nr. of Retweets	210	418
Mean Nr. of Likes	752	2066

Figure C5

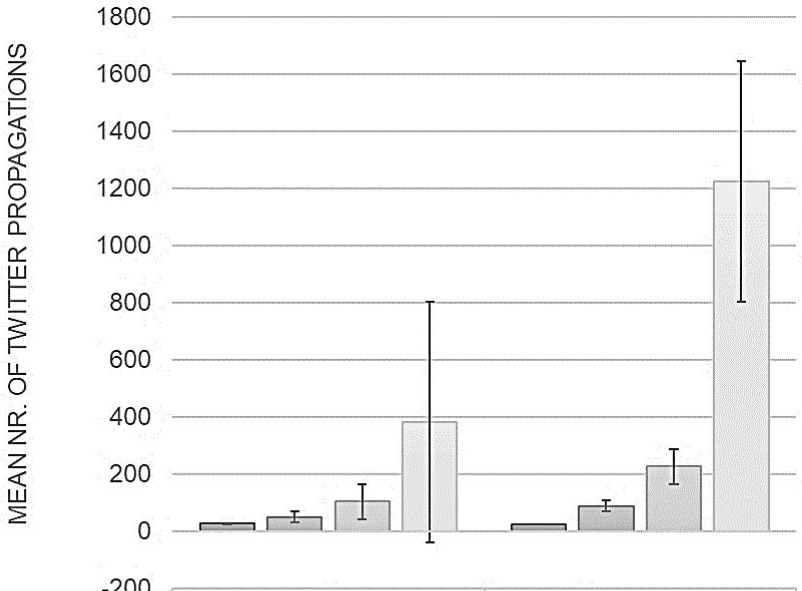
Politicians: Numbers of Tweets per Binary Humour-Degree



	Serious	Humorous
Nr. of Tweets	200051	25041
Proportion of Tweets	88.88%	11.12%

Figure C6

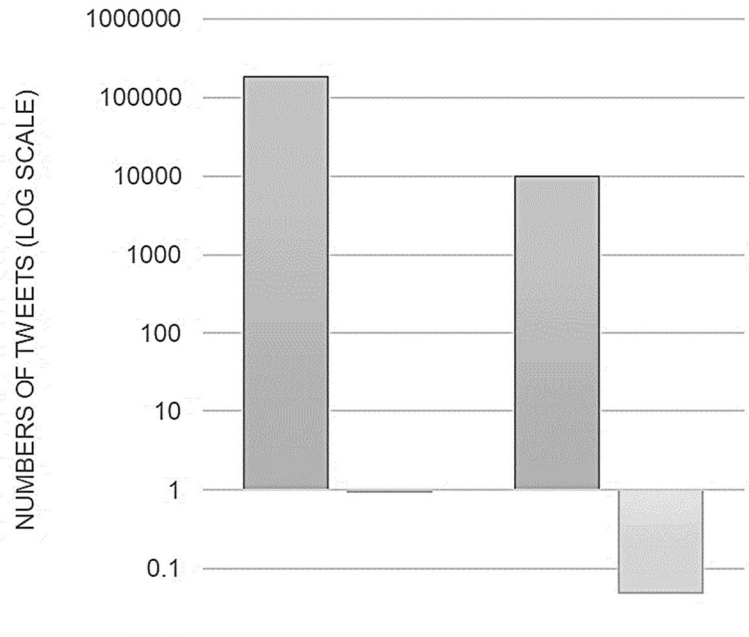
Political Journalists: Mean Propagations per Binary Humour-Degree (SE of Mean)



	Serious	Humorous
Mean Nr. of Quotes	26	25
Mean Nr. of Replies	49	88
Mean Nr. of Retweets	104	227
Mean Nr. of Likes	382	1225

Figure C7

Political Journalists: Numbers of Tweets per Binary Humour-Degree



	Serious	Humorous
Nr. of Tweets	185098	9833
Proportion of Tweets	94.96%	5.04%

Figure C8

Comedians: Mean Propagations per Binary Humour-Degree (SE of Mean)

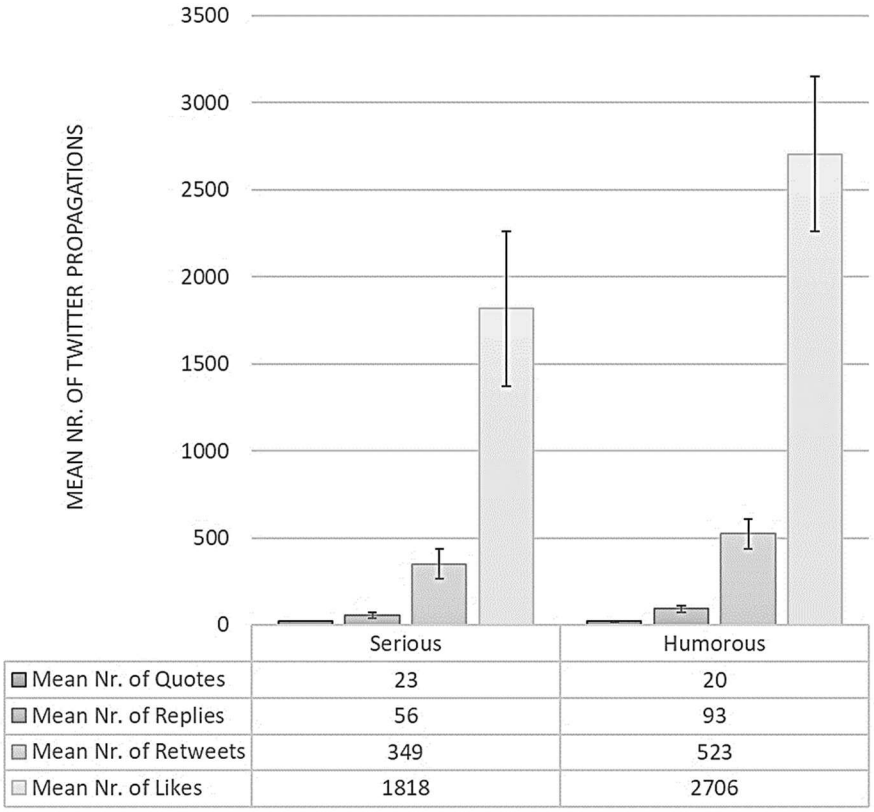
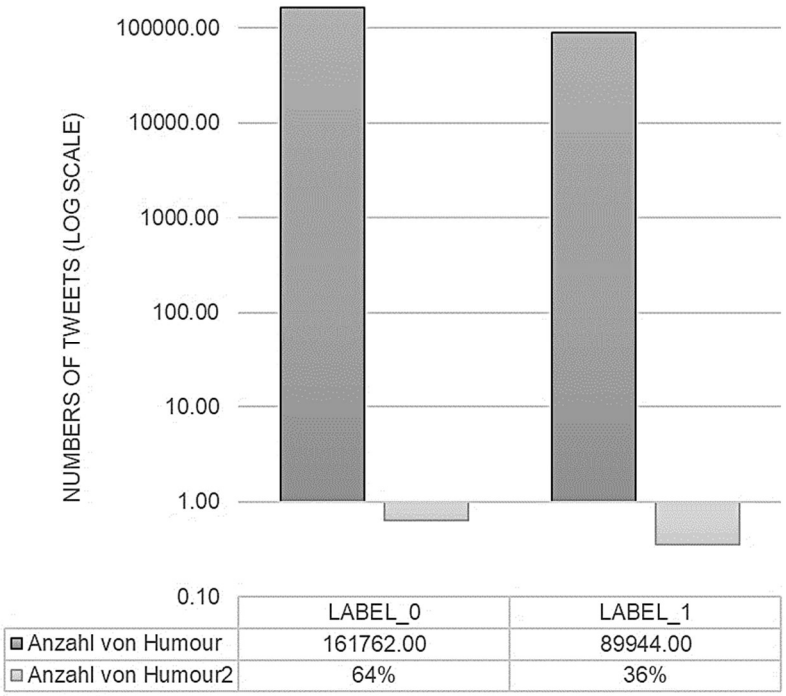


Figure C9

Comedians: Numbers of Tweets per Binary Humour-Degree

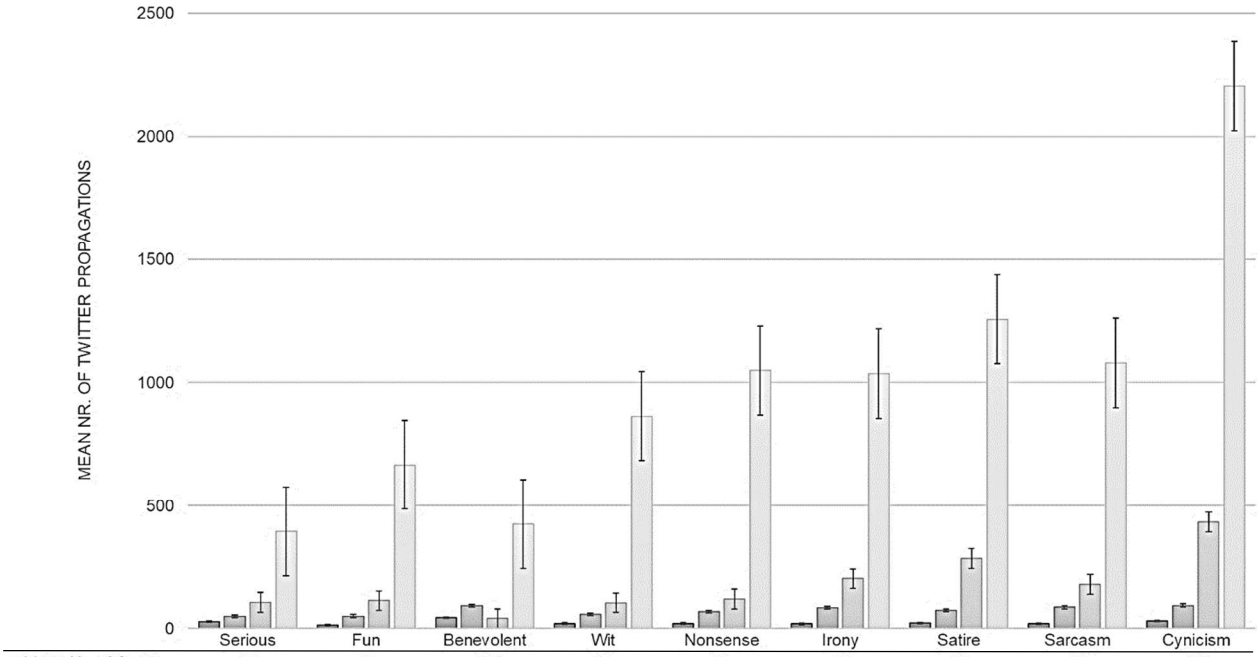


Appendix D

Correlation of Humour in Political Tweets to Virality Metrics: Political Journalists

Figure D1

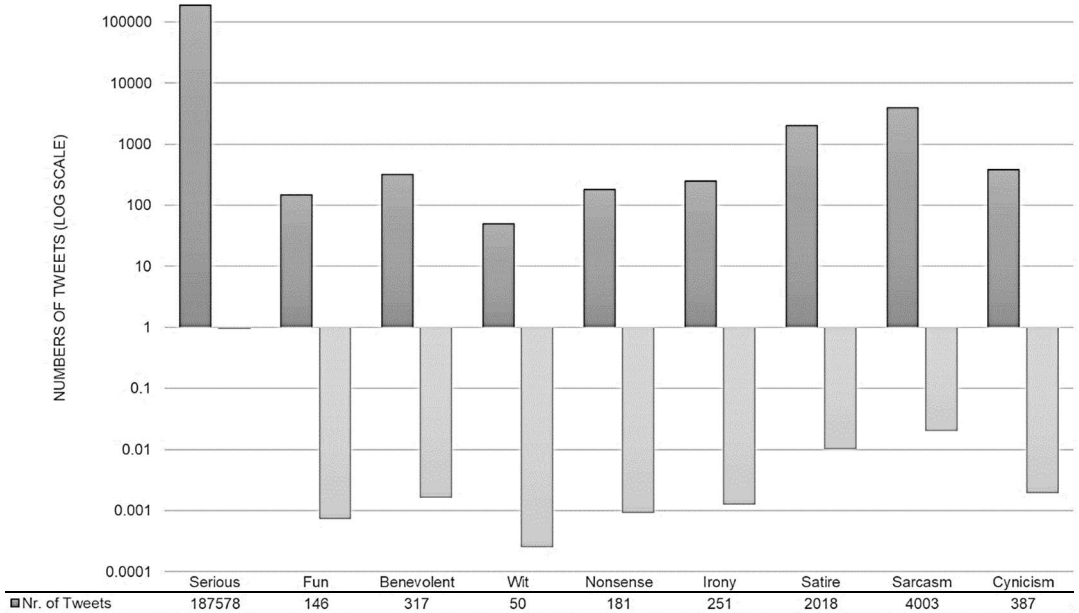
Political Journalists: Mean Propagations per Humour-type (SE of Mean)



	Serious	Fun	Benevolent	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Mean Nr. of Quotes	27	15	44	21	21	18	22	19	30
Mean Nr. of Replies	50	51	93	57	68	85	74	86	94
Mean Nr. of Retweets	106	113	40	105	120	202	285	179	433
Mean Nr. of Likes	396	667	425	864	1049	1036	1257	1080	2204

Figure D2

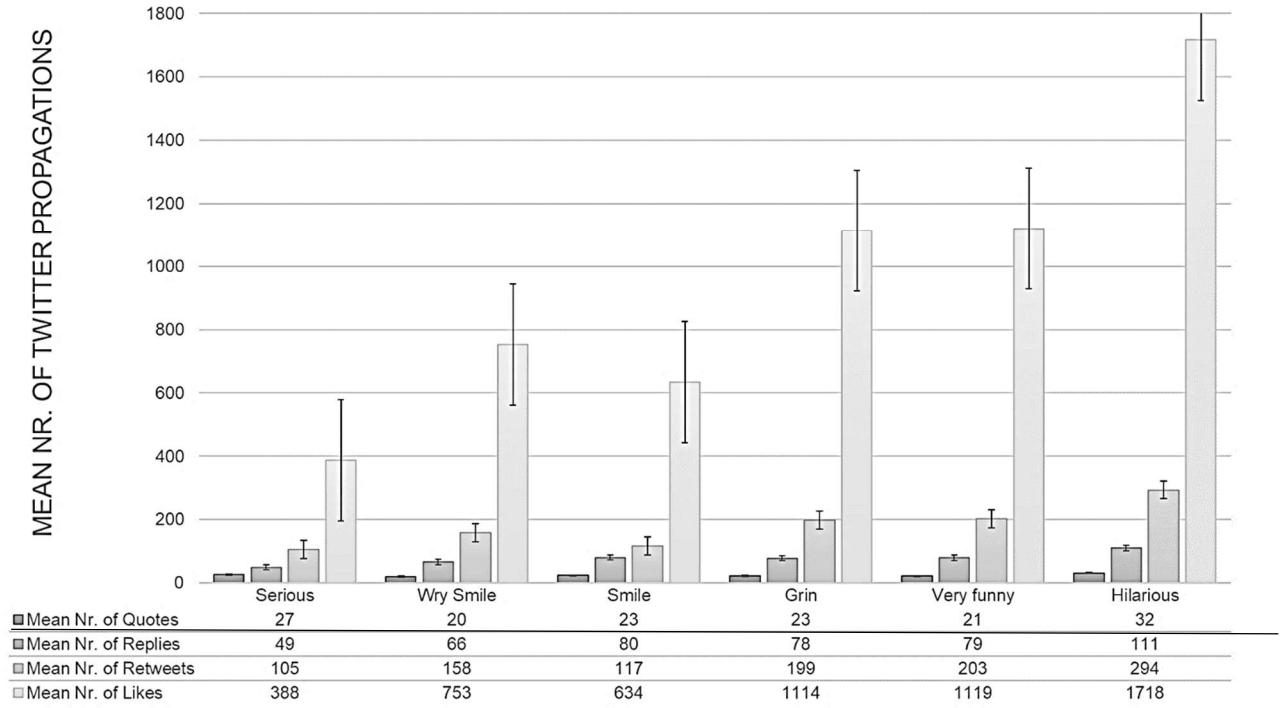
Political Journalists: Numbers of Tweets per Humour-type



	Serious	Fun	Benevolent	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Nr. of Tweets	187578	146	317	50	181	251	2018	4003	387
Proportion of Tweets	96.23%	0.07%	0.16%	0.03%	0.09%	0.13%	1.04%	2.05%	0.20%

Figure D3

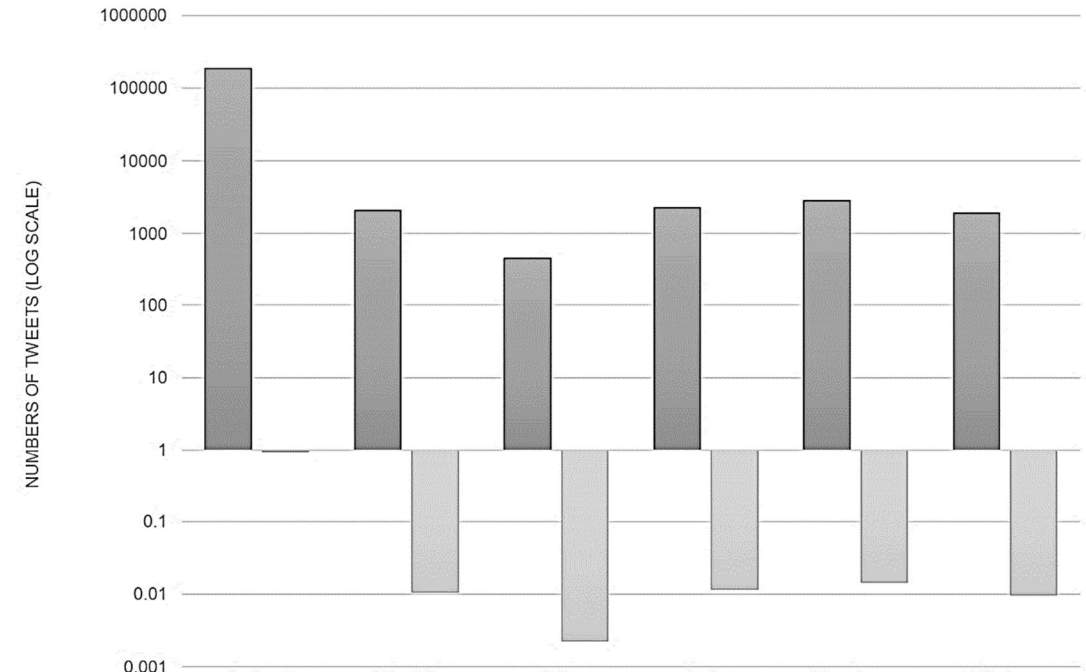
Political Journalists: Mean Propagations per Humour-Degree (SE of Mean)



	Serious	Wry Smile	Smile	Grin	Very funny	Hilarious
Mean Nr. of Quotes	27	20	23	23	21	32
Mean Nr. of Replies	49	66	80	78	79	111
Mean Nr. of Retweets	105	158	117	199	203	294
Mean Nr. of Likes	388	753	634	1114	1119	1718

Figure D4

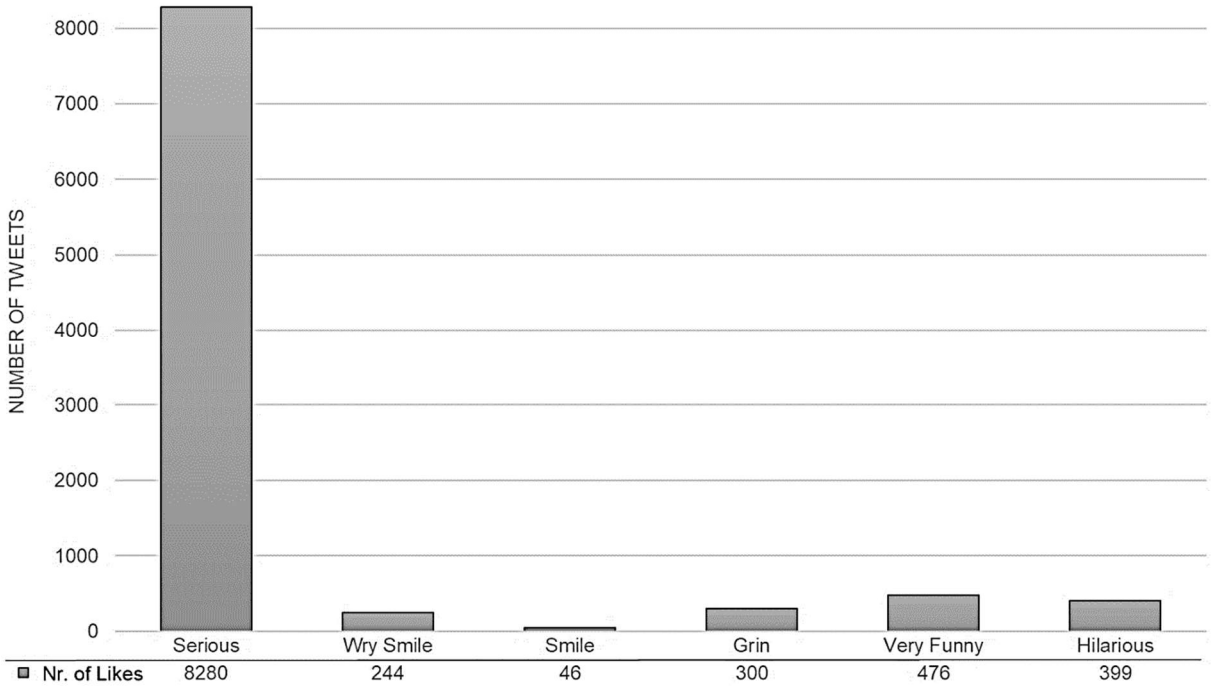
Political Journalists: Numbers of Tweets per Humour-Degree



	Serious	Wry grin	Grin	Smile	Very funny	Hilarious
Nr. of Tweets	185367	2075	448	2274	2873	1894
Proportion of Tweets	95.09%	1.06%	0.23%	1.17%	1.47%	0.97%

Figure D5

Political Journalists: Distribution of Upper 5% of Liked Tweets

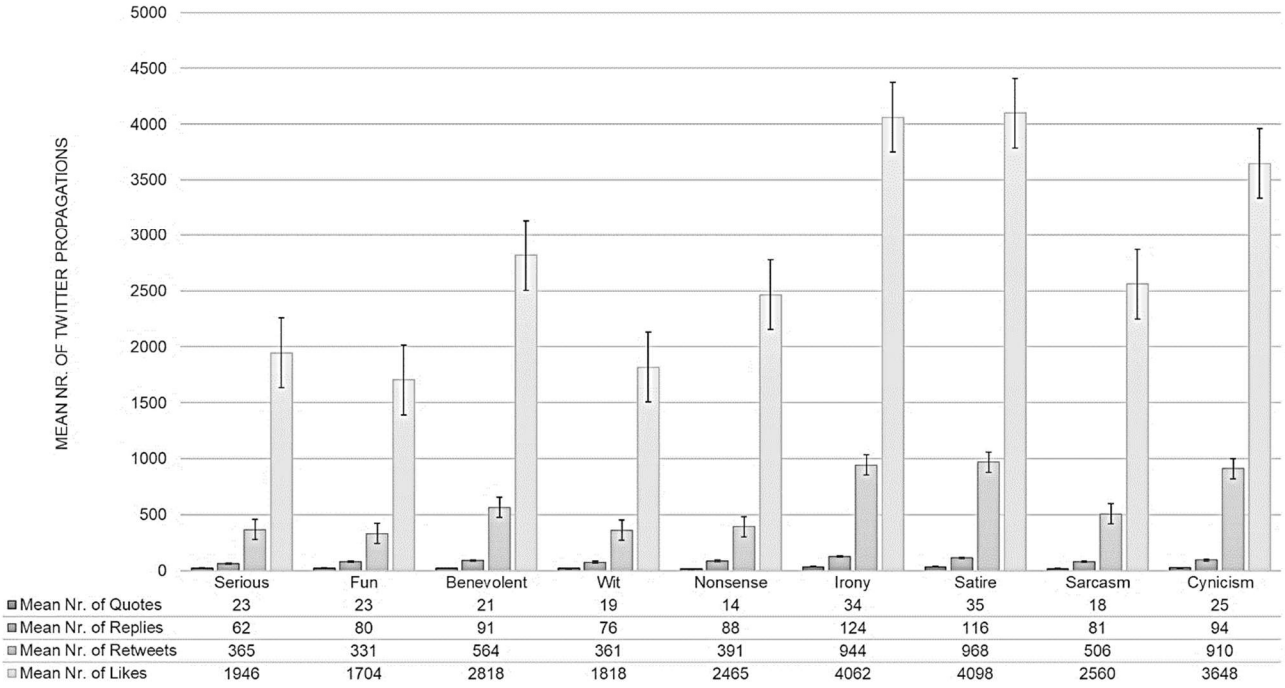


Appendix E

Correlation of Humour in Political Tweets to Virality Metrics: Comedians

Figure E1

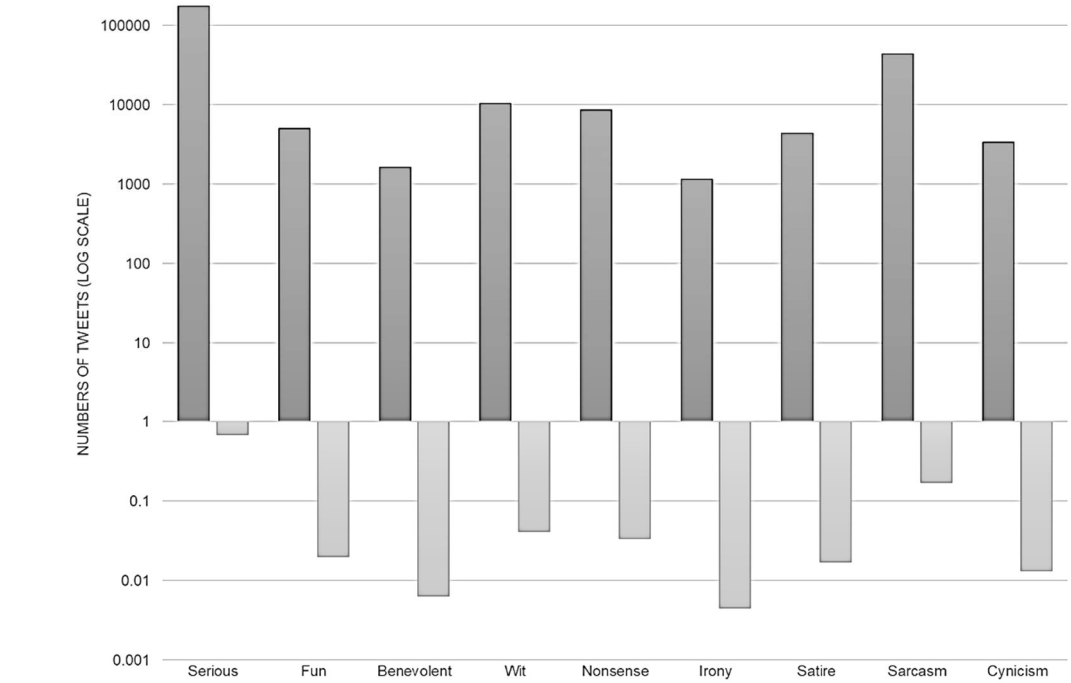
Comedians: Mean Propagations per Humour-Type (SE of Mean)



	Serious	Fun	Benevolent	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Mean Nr. of Quotes	23	23	21	19	14	34	35	18	25
Mean Nr. of Replies	62	80	91	76	88	124	116	81	94
Mean Nr. of Retweets	365	331	564	361	391	944	968	506	910
Mean Nr. of Likes	1946	1704	2818	1818	2465	4062	4098	2560	3648

Figure E2

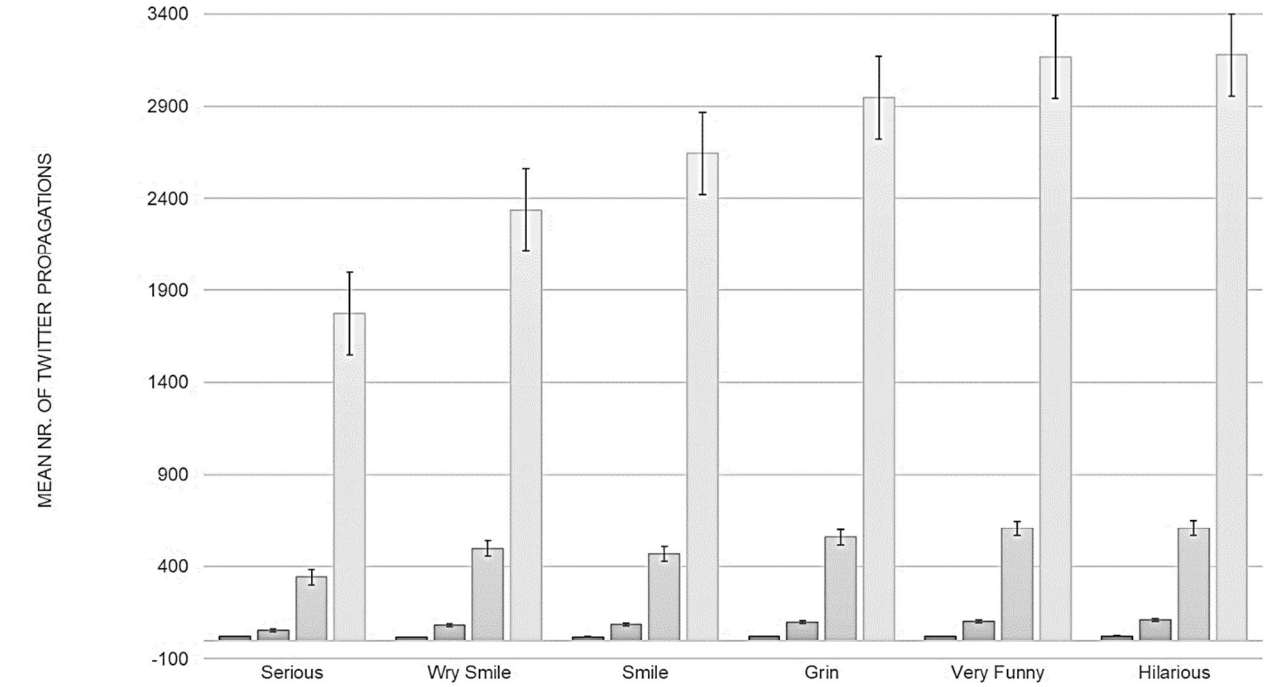
Comedians: Numbers of Tweets per Humour-Type



	Serious	Fun	Benevolent	Wit	Nonsense	Irony	Satire	Sarcasm	Cynicism
Nr. of Tweets	173612	5034	1630	10439	8554	1151	4361	43571	3354
Proportion of Tweets	68.97%	2.00%	0.65%	4.15%	3.40%	0.46%	1.73%	17.31%	1.33%

Figure E3

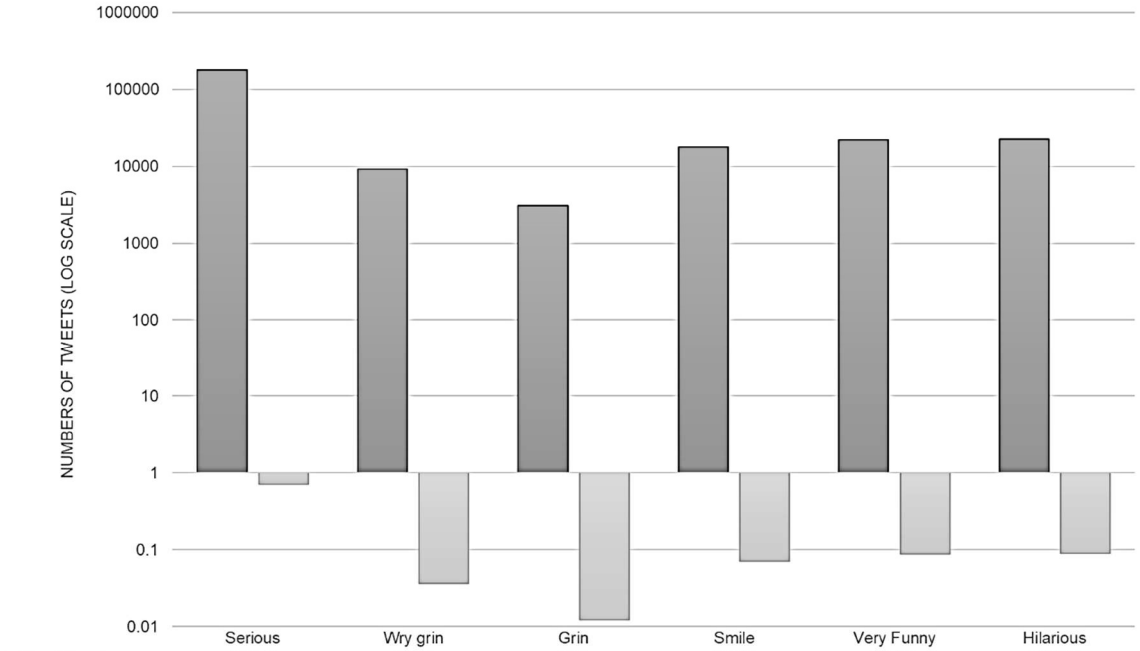
Comedians: Mean Propagations per Humour-Degree (SE of Mean)



	Serious	Wry Smile	Smile	Grin	Very Funny	Hilarious
Mean Nr. of Quotes	22	19	19	21	23	23
Mean Nr. of Replies	56	81	87	98	101	112
Mean Nr. of Retweets	342	495	467	557	607	608
Mean Nr. of Likes	1772	2340	2644	2946	3167	3178

Figure E4

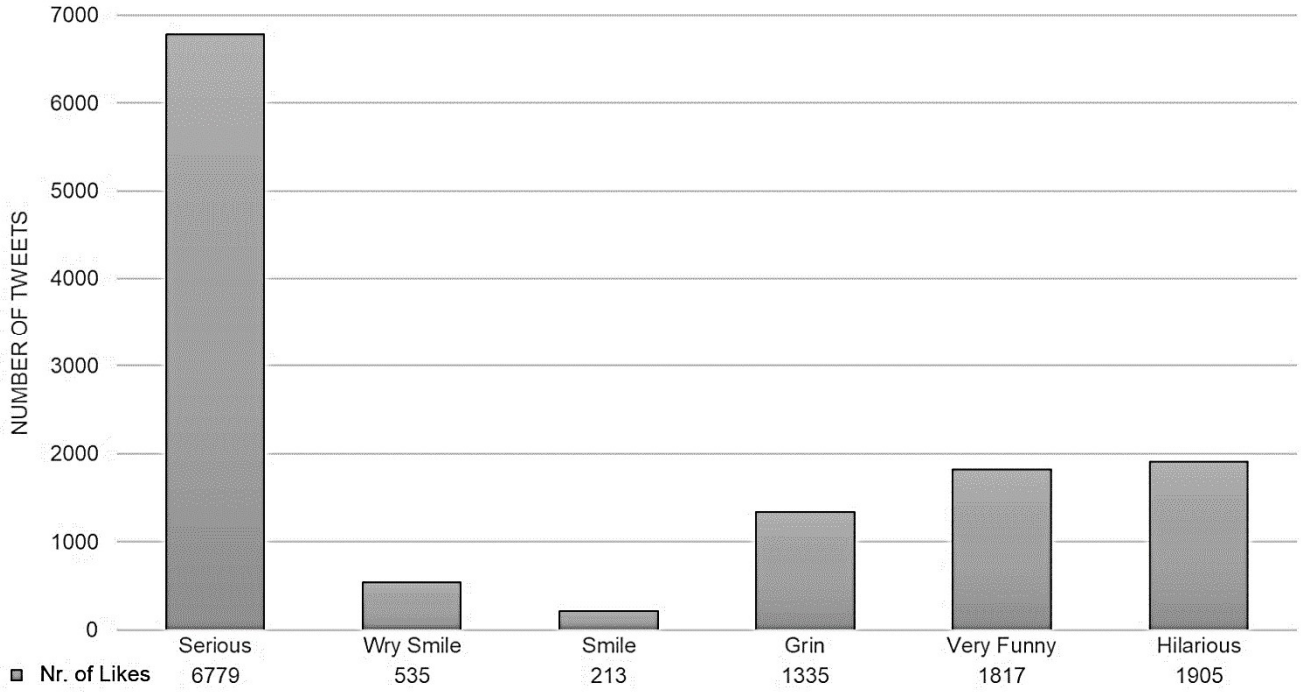
Comedians: Numbers of Tweets per Degree of Humour



	Serious	Wry grin	Grin	Smile	Very Funny	Hilarious
Nr. of Tweets	176993	9102	3068	17809	22182	22552
Proportion of Tweets	70.32%	3.62%	1.22%	7.08%	8.81%	8.96%

Figure E5

Comedians: Distribution of Upper 5% of Liked Tweets



Appendix F

Twitter Handles Used in Data Procurement Phase

UK Members of Parliament

@BorisJohnson, @jeremycorbyn, @Keir_Starmer, @theresa_may, @Ed_Miliband, @DavidLammy, @CarolineLucas, @RishiSunak, @Jacob_Rees_Mogg, @jessphillips, @MattHancock, @HackneyAbbott, @pritipatel, @DominicRaab, @YvetteCooperMP, @johnmcdonnellMP, @timfarron, @AngelaRayner, @Jeremy_Hunt, @michaelgove, @EmilyThornberry, @sajidjavid, @lisanandy, @MhairiBlack, @DrRosena, @RLong_Bailey, @HarrietHarman, @DavidDavisMP, @hilarybennmp, @TulipSiddiq, @DawnButlerBrent, @zarahsultana, @stellacreasy, @RichardBurgon, @grantshapps, @andrealeadsom, @trussliz, @joannaccherry, @SteveBakerHW, @Ianblackford_MP, @labourlewis, @LiamFox, @RachelReevesMP, @johnredwood, @GavinWilliamson, @JonAshworth, @JamesCleverly, @leicesterliz, @BarryGardiner, @JohnnyMercerUK, @TanDhesi, @IanLaveryMP, @angelaeagle, @wesstreeting, @PennyMordaunt, @NadiaWhittomeMP, @BenPBradshaw, @DanJarvisMP, @RhonddaBryant, @LaylaMoran, @EdwardJDavey, @AnnelieseDodds, @EstherMcVey1, @andreajenkyns, @Tobias_Ellwood, @TomTugendhat, @nadhimzahawi, @BrandonLewis, @LucyMPowell, @TracyBrabin, @Alison_McGovern, @RobertJenrick, @Debbie_abrahams, @MPIainDS, @OwenPaterson, @KateGreenSU, @PeteWishart, @GwynneMP, @BellRibeiroAddy, @margarethodge, @DanCardenMP, @NazShahBfd, @jreynoldsMP, @meaglemp, @SarahChampionMP, @LouHaigh, @AlokSharma_RDG, @RupaHuq, @AlynSmith, @ClaudiaWebbe, @ChiOnwurah, @DesmondSwayne, @rushanaraali, @jon_trickett, @KemiBadenoch, @AndyMcDonaldMP, @Geoffrey_Cox, @AngusMacNeilSNP, @CatSmithMP, @Mike_Fabricant, @peterkyle, @MarshadeCordova, @PeterBoneUK, @LindsayHoyle_MP, @bphillipsonMP, @tracey_crouch, @JohnPenroseNews, @GregHands, @Dr_PhilippaW, @liambyrnemp, @columeastwood, @bernardjenkin, @DehennaDavison, @ABridgen, @TommySheppard, @PreetKGillMP, @KateOsamor, @halfon4harlowMP, @LukePollard, @NeilDotObrien, @SDoughtyMP, @RosieDuffield1, @SharonHodgsonMP, @SKinnock, @ClaireHanna, @SteveReedMP, @KarlTurnerMP, @BBradley_Mans, @StewartMcDonald, @DamianHinds, @Douglas4Moray, @JoStevensLabour, @JulianSmithUK, @MrJohnNicolson, @IanMurrayMP, @SteveBarclay, @KerryMP, @OliverDowden, @grahamemorris, @GregClarkMP, @Bill_Esterson, @SuellaBraverman, @CatherineWest1, @DamianCollins, @lloyd_rm,

@NickTorfaen, @JBrokenshire, @CharlesWalkerMP, @JohnHealey_MP, @Mark_J_Harper, @LilianGreenwood, @johnfinucane, @lucyallan, @RachaelMaskell, @darrenpjones, @alisonthewliss, @ThangamMP, @CatMcKinnell, @StewartHosieSNP, @GeorgeFreemanMP, @IanMearnsMP, @RobertBuckland, @Afzal4Gorton, @EmmaHardyMP, @hbaldwin, @tobyperkinsmp, @DianaJohnsonMP, @JamesDuddridge, @IanByrneMP, @Jesse_Norman, @coyleneil, @KeeleyMP, @justinmadders, @theresecoffey, @JimfromOldham, @hammersmithandy, @KirstySNP, @DKShrewsbury, @carolynharris24, @BobBlackman, @SeemaMalhotra1, @FloEshalomi, @KevinBrennanMP, @Andrew4Pendle, @JackDromeyMP, @EmmaLewellBuck, @HannahB4LiviMP, @patmcfaddenmp, @Helen_Whately, @HollyLynch5, @ConorBurnsUK, @alexsobel, @helenhayes_, @Nus_Ghani, @ApsanaBegumMP, @SCrabbPembs, @KevanJonesMP, @JohnGlenUK, @CSkidmoreUK, @carolinenokes, @ShabanaMahmood, @annietrev, @RuthCadbury, @YasminQureshiMP, @elliereeves, @chhcalling, @BarrySheerman, @PaulBlomfieldMP, @charlotte2153, @cj_dinenage, @MGreenwoodWW, @ACunninghamMP, @mariacaulfield, @mtpennycook, @SamTarry, @J_Donaldson_MP, @MariaMillerUK, @AndrewRosindell, @amcarmichaelMP, @BillCashMP, @Wera_Hobhouse, @PhilipDaviesUK, @JulieElliottMP, @AnneMarieMorris, @JonCruddas_1, @BWallaceMP, @LabourSJ, @JustinTomlinson, @eastantrimmp, @AndrewSelous, @DanielZeichner, @ChrisHazzardSF, @vickyfoxcroft, @KwasiKwarteng, @Siobhain_Mc, @ChrisPincher, @vickyford, @CMonaghanSNP, @JakeBerry, @DamianGreen, @Meg_HillierMP, @timloughton, @Pauline_Latham, @KennyMacAskill, @HeatherWheeler, @gildernewm, @HelenGrantMP, @VirendraSharma, @S_Hammond, @GuyOpperman, @mattwarman, @ChrisLawSNP, @JDjanogly, @StuartAndrew, @libdemdaisy, @PaulMaskeyMP, @PaulMaynardUK, @Rehman_Chishti, @HuwMerriman, @Imran_HussainMP, @sheryllmurray, @JGray, @michelledonelan, @NickGibbUK, @NorwichChloe, @steve_mccabe, @iainastewart, @SarahOwen_, @Mark_Spencer, @NiaGriffithMP, @JasonMcCartney, @stephenctimms, @KarenPBuckMP, @ConorMcGinn, @TaiwoOwatemi, @LSRPlaid, @SirGrahamBrady, @Simon4NDorset, @neill_bob, @HenrySmithUK, @kitmalthouse, @grahamstuart, @SimonClarkeMP, @DeidreBrock, @GradySNP, @DavidMundellDCT, @willquince, @MargaretFerrier, @PeterGrantMP, @GarethThomasMP, @AdamAfriyie, @ChrisStephens, @sarahjolney1, @AnnaMcMorrin, @HuddlestonNigel, @mimsdavies, @AlecShelbrooke, @Peter_Dowd, @MarcusFysh, @DavidTCDavies, @Mark4WyreForest, @drewhendrySNP, @AJonesMP, @abenaopp, @AWMurrison, @joymorrissey, @kirstenoswald, @AndrewBowie_MP,

@KellyTolhurst, @GeraintDaviesMP, @scullyp, @neilgraysnp, @BrendanOHaraMP, @morton_wendy, @DJWarburton, @amandamilling, @JudithCummins, @ToniaAntoniazzi, @DougChapmanSNP, @JSHeappey, @ChrisM4Chester, @SMcPartland, @stevedouble, @WalkerWorcester, @GavNewlandsSNP, @William_Wragg, @WayneDavid_MP, @AmyCallaghanSNP, @aliciakearns, @davidmorrisml, @CherylGillan, @CliveEfford, @BimAfolami, @DavidEvennett, @DrLisaCameronMP, @AlunCairns, @_OliviaBlake, @KimJohnsonMP, @GRobinsonDUP, @cajardineMP, @rosie4westlancs, @DavidJonesMP, @Y_FovargueMP, @nigelmp, @PaulaBarkerMP, @MartinJDocherty, @lynbrownmp, @AnneMcLaughlin, @Michael_Ellis1, @pow_rebecca, @jessicamordenmp, @John4Carlisle, @DavidRutley, @PGibsonSNP, @Stuart_McDonald, @ronniecowan, @nadams, @MarieRimmer, @munirawilson, @CPhilpOfficial, @AlexChalkChelt, @MartinVickers, @VotePursglove, @StephenFarryMP, @IoWBobSeely, @EdwardLeighMP, @StephenMorganMP, @CGreenUK, @JackieDP, @RichardGrahamUK, @marykfoy, @VictoriaPrentis, @Jochurchill4, @AlanBrownSNP, @karinsmyth, @Jamie4North, @GillianKeegan, @DavidLinden, @julianknight15, @marionfellows, @kevinhollinrake, @RobertSyms, @PutneyFleur, @John2Win, @OliverHealdUK, @Dunne4Ludlow, @CrispinBlunt, @KateOsborneMP, @EddieHughes4WN, @Rees4Neath, @cmackinlay, @MikeKaneMP, @GregKnight, @peter_aldous, @JanetDaby, @BrineMP, @MickWhitleyMP, @MartynDaySNP, @karlmccartney, @NavPMishra, @spellar, @JeffSmithetc, @redditchrachel, @MikeAmesburyMP, @MPritchardUK, @CPJElmore, @danny__kruger, @AlbertoCostaMP, @Michael4MDNP, @FeryalClark, @Siobhan_Baillie, @HywelPlaidCymru, @AlanMakMP, @Lee4NED, @DarrenG_Henry, @Offord4Hendon, @BlaenauGwentMP, @Steph_Peacock, @Simonhartmp, @CWhittaker_MP, @jc4southsuffolk, @Bambos_MP, @JacobYoungMP, @Marcus4Nuneaton, @MarkTamiMP, @robertcourts, @MaryRobinson01, @JohnMcNallySNP, @tomhunt1988, @eleanor4epping, @craig4nwarks, @MattWestern_, @tony4rochdale, @lucyfrazermp, @paulbristow79, @Kate_HollernMP, @carlalockhart, @JamesMorris, @JWhittingdale, @scottmann4NC, @jamesmurray_ldn, @BethWinterMP, @JackLopresti, @drlukeevans, @alanwhiteheadmp, @BenMLake, @RicHolden, @Valerie_VazMP, @AlexDaviesJones, @Royston_Smith, @Nicola4WBE, @OrfhlaithBegley, @kevin_j_foster, @FrancieMolloy, @JHowellUK, @rach_hopkins, @OwenThompson, @MarkPawsey, @AlexNorrisNN, @AJRichardsonMP, @theodoraclarke, @DerekTwiggMP, @ClaireCoutinho, @ElliotColburn, @Ben_Everitt, @GarethDavies_MP, @bhatti_saqib, @Metcalfe_SBET, @GeraldJonesLAB,

@pauljholmes, @AthertonNWales, @RebeccaHarrisMP, @mikejwood, @Chris_EvansMP, @MattRodda, @gregsmith_uk, @neil_parish, @Stuart4WolvesSW, @GillFurnissMP, @edwardtimpson, @Laura__Farris, @FabianLeedsNE, @alancampbellmp, @PaulGirvanMP, @LeoDochertyUK, @SirRogerGale, @drcarolinej, @JuliaLopezMP, @MelJStride, @imranahmadkhan, @david_duguid, @robertlargan, @Caroline_Ansell, @JonesyFay, @RuthNewportWest, @FelicityBuchan, @khalid4PB, @JamesDavies, @JohnCryerMP, @wendychambLD, @Matt_VickersMP, @simonjamesjupp, @chrisloder, @GarethBaconMP, @NatalieElphicke, @ChrisClarksonMP, @craig4monty, @JNHanvey, @nigelmills, @Alex_Stafford, @DerekThomasUK, @StephenFlynnSNP, @AnthonyMangnall1, @MickeyBradySF, @MrAndy_Carter, @SarBritcliffeMP, @MpHendrick, @LukeHall, @ASollowayUK, @ALewerMBE, @AaronBell4NUL, @garystreeterSWD, @_RobbieMoore, @amessd_southend, @DrBenSpencer, @JamesDalyMP, @Christian4BuryS, @mark4dewsbury, @Jane_Stevenson_, @GilesWatling, @GarySambrook89, @baynes_simon, @ScottBentonMP, @markjenkinsonmp, @alexburghart, @ShaileshVara, @LizTwistMP, @ab4scambs, @lrobertsonTewks, @JulieMarsonMP, @Q66Suzi, @griffitha, @RuthEdwardsMP, @KieranMullanUK, @simonfell, @gaganmohindra, @jogideon, @DaveDooganSNP, @SelaineSaxby, @Tom_Randall, @Shaun4WBW, @StevenBonnarSNP, @marcolonghi4dn, @K_Fletcher_MP, @MikeHillMP, @antony_hig, @SallyAnn1066, @dean4watford, @DSimmonds_RNP, @lia_nici, @DrNeilHudson, @duncancbaker, @JaneMHunt, @JamesSunderl, @RobinMillarMP, @TahirAliMP, @RThomsonMP, @JimShannonMP, @Gibbo4Darlo, @Bren4Bassetlaw, @NickFletcherMP, @YasinForBedford, @jamesowild, @PaulHowellMP, @twocitiesnickie, @JamieWallisMP, @thisischerilyn, @JonathanLord, @david4wantage, @JeromeMayhew, @TeamRanil, @JulianSturdy, @Dines4Dales, @richardbaconmp, @maggie_erewash, @AylesburyTories, @GlindonMary, @LeeAndersonMP

Political Journalists

@10downingstreet, @AJEnglish, @AldeburghCinema, @AlfieConn, @AliceBhand, @AnnaAdamsBBC, @Anna_Filip, @BBCAllegra, @BBCBreakfast, @BBCDomC, @BBCJonSopel, @BBCLN, @BBCLookNorth, @BBCMarkDenten, @BBCNews, @BBCNewsnight, @BBCParliament, @BBCPolitics, @BBCRadio2, @BBCRadio4, @BBCRichardMoss, @BBCTheOneShow, @Barnes_Runners, @C4election, @Channel4News, @CharlieBeckett, @ChrisMasonBBC, @ChrisWimpress, @ColinHazelden, @ConHome, @Conhome, @DAaronovitch, @DWNews,

@DailyMailUK, @DailyMirror, @Daily_Record, @DavidPBMaddox, @DavidWooding, @David_Stringer, @E_D_Dagan, @EdConwaySky, @FT, @ForTheManyPod, @ForesightNewsUK, @FraserNelson, @Freedland, @GMB, @GaryGibbonC4, @GdnPolitics, @GeordieStory, @GerriPeev, @GillPenlington, @GuidoFawkes, @IainDale, @IainDaleAllTalk, @IanMurrayHall, @IndyPolitics, @JBeattieMirror, @JEagleshamFT, @JEvansTimes, @JPonpolitics, @James_Macintyre, @JewishChron, @JohnRentoul, @KathViner, @Kevin_Maguire, @LBC, @LizzyBuchan, @LornyDunkley, @MSNBC, @MarkAnsell, @MattChorley, @MirandaUK, @MirrorPolitics, @MishalHusain, @Morkins, @NFFC, @NYTimes, @NatalieHanman, @NewStatesman, @PA, @PaulGoodmanCH, @PaulLewis, @PeterSpencer, @PippaCrerar, @RaynerSkyNews, @ReactionLife, @RobDotHutton, @SamCoatesSky, @SamiraAhmedUK, @SaraMojtehedz, @SimonJeffery, @SkyNews, @SkyNewsBreak, @StAlbansCityFC, @StephenNolan, @TheTimes, @Time, @Torcuil, @TorontoStar, @UKParliament, @VictoriaPeckham, @WP_Radio, @WorldVision, @WorldVisionUK, @Zain_Verjee, @_alexforrest, @_fionawalsh, @adamboultonSKY, @adavies4, @alexmassie, @alextomo, @alicecpark, @allegrastratton, @amolrajan, @andrewrawnsley, @andybell5news, @anntreneman, @artemisq, @bbccumbria, @bbclaurak, @bbcnewcastle, @bbcnews, @bbcnewsnight, @bbcradio4, @bbctees, @beisgovuk, @ben_duckworth, @benedictbrogan, @benglaze, @benm_d, @bhamlaw, @bonisones2, @bp_plc, @brittanyvgreer, @business, @bysshe1, @c4marcus, @camanpour, @cathynewman, @channel4news, @cherylsmith, @chrishams, @chrisshipitv, @christian_aid, @christopherhope, @climate, @cnni, @craigawoodhouse, @daily_politics, @danbloom1, @davidallengreen, @davidcornock, @dickiedogsfc, @duchenneuk, @eyespymp, @faisalislam, @gabyhinsliff, @gbnews, @gemma_charles, @glenoglaza1, @grahamdines, @guardian, @hyper_drive, @iainmartin1, @iankatz1000, @idanha_simon, @insidepmi, @itvnews, @janemerrick23, @jasonmillspmi, @jenlipman, @jerryhayes1, @jmackin2, @joeyfjones, @johannhari101, @joncraig, @jonsnowC4, @kirstywalker1, @law_and_policy, @lindseyhilsum, @lnmulholland, @lucymanning, @martinkettle, @mattwardman, @mccannlondon, @mehdihasanshow, @mehdirhasan, @mikeysmith, @moneysavingexp, @nicholaswatt, @oliverjamesking, @openDemocracy, @parkrunUK, @patrickwintour, @peacockTV, @pghoskin, @politicseditor, @pollycurtis, @prospect_clark, @prospect_uk, @reactionlife, @rosschawkins, @seandilleyNEWS, @sgreilly, @shanegreer, @sheffielduni, @simonsketch, @skygillian, @skynewsniall, @sophieelmhirst, @stephenpollard, @stuartmillar159, @sunderlanduni,

@sunny_hundal, @telegraph, @theDefaultLine, @theipaper, @thepcrc, @theterminal, @thetimes, @timespolitics, @timesradio, @timesredbox, @toby_etc, @tobyhelm, @tombod, @tortoise, @tribunemagazine, @vincentmoss, @whwoodward, @xtophercook

Serious News

@AP, @BBCworld, @ITN, @ITVnews, @SkyNewsPolitics, @TheEconomist

Comedians

@bechillcomedian, @robdelaney, @jimaffigan, @meganamram, @susancalman, @juliussharpe, @adambuxton, @weemissbea, @almurray, @alancarr, @alandavies1, @alex_brooker, @billbailey, @charltonbrooker, @mrchrisaddison, @IAmChrisRamsey, @daraobriain, @davegorman, @baddiel, @RealDMitchell, @davidschneider, @davidwalliams, @MrEdByrne, @eddieizzard, @frankieboyle, @grahnort, @gdavies, @therealjackdee, @jackwhitehall, @JKCorden, @jasonmanford, @jimmycarr, @joelycett, @gillinghamjoe, @joeldommett, @JohnBishop121, @johncleese, @JohnnyVegasReal, @ronjichardson, @wossy, @jonnyawsum, @joshwiddicombe, @kerrygodliman, @mslouatkinson, @realmatbaynton, @realmattlucas, @McInTweet, @themiltonjones, @mermhart, @MoTheComedian, @mrnishkumar, @noelfielding10, @richardayoade, @rickygervais, @robbeckettcomic, @RobBrydon, @arobertwebb, @robinince, @roisinconaty, @romeshranga, @realrossnoble, @rufushound, @rustyrockets, @russellhoward, @sarapascoe, @SarahMillican95, @seannwalsh, @stephenfry, @stephenmangan, @sueperkins, @timminchin, @RealTimVine, @tomcraine, @badbanana, @stevemartingtogo, @conanobrien, @stephenathome, @theonion, @preschoolgems, @boredelonmusk, @thetweetofgod, @mindykaling, @haveigotnews, @therealjackdee, @mattforde, @kevinbridges86, @jeremyclarkson, @roisinconaty, @sueperkins, @richardosman, @kathbum, @stephencgrant

Appendix G

Recommendations for Hyperparameter Optimization

In this appendix some advice for other researchers is summarized as a result of learnings from the lengthy phase of fine-tuning NLMs in this project.

Training Epochs

Variations in number of training epochs (between 1 and 50) for the same volume of training data had in most cases a surprisingly limited effect. Typically 3 to 5 epochs is sufficient, but this varies a little according to the model. The medium and large Ernie models proved able to complete a fairly competitive fine-tuning after just 2 or even 1 training-epoch.

Performance

The use of GPUs over equivalent CPUs is always to be recommended during fine-tuning but also in the inference phase. A performance increase of 5x was typically experienced.

If able to utilise a GPU with an appropriate driver, then FP16 (16-bit floating point precision) is a huge boost to performance during training was observed and FP16 was thus used extensively throughout to project, giving a performance improvement of roughly 2x over "full fp".

Models (with hyperparameter combinations) which are faster to train are also faster when being used for inference (see figure G1).

Batch size should be maximised to improve performance, but this is limited by the amount of RAM available combined with the size of model. An optimum can be established with a very short test and can potentially save many hours of training. For instance, the GPT2-Large model could only be trained with a batch size of 1, due to memory constraints (GPU with 16GB RAM).

The interactive visualizations available from W&B – in particular the “parallel coordinates” and “parameter importance” charts – helped in finding which models should be chosen as HDS for each case (see figures 1 and 5 for examples).

Figure G1

Training Performance: Samples/s (Training vs. Evaluation)

Note: Scatterplot taken from: [Project Dashboard (wandb.ai)]

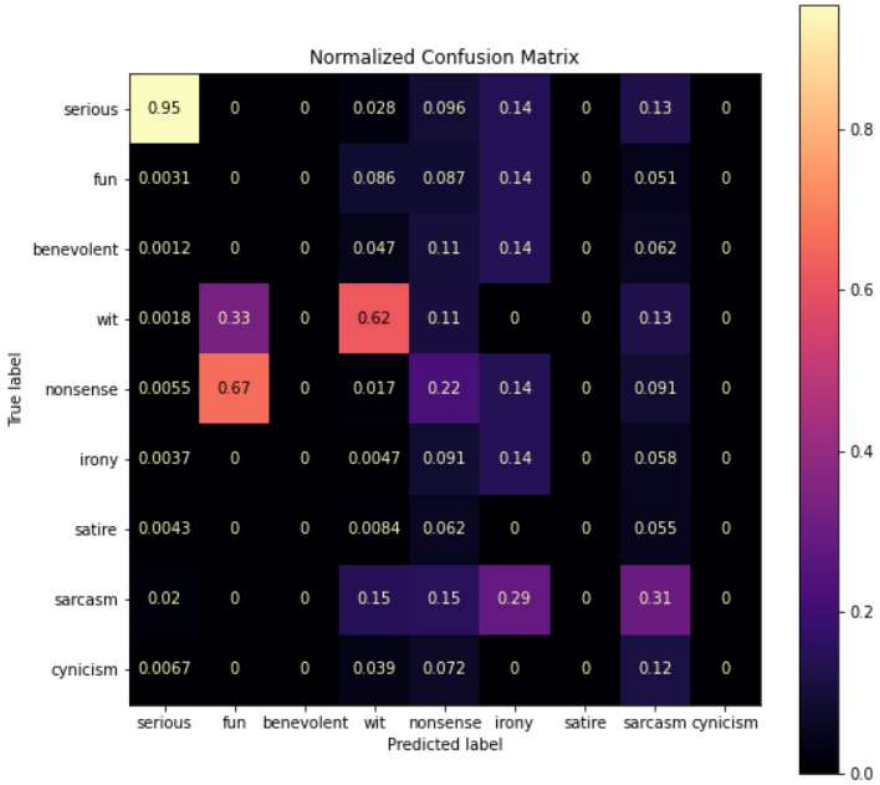
Model Size

The distilled versions (DistilBERT, DistilRoBERTa and to a lesser extent DistilGPT2) appear to provide a performance boost of up to 100%. Additionally, smaller models have the advantage that due to their smaller footprint, more RAM is available and allows for large batch-sizes, further increasing training performance.

Special Case: Small NLMs. Of the AE architectures it was noticeable that the smallest model used: Ernie-tiny – a compressed model from Ernie 2.0 utilising model structure compression and model distillation for performance – was completely unable to detect 4 of the humour-types (figure G2), with both 3 and 5 training epochs (see discussion in Limitations and Future Research).

Figure G2

Normalized Confusion Matrix of Ernie-tiny (5 training epochs)



Other than the GPT2 derivatives and the Ernie-tiny models, all others showed very balanced values for Precision and Recall and produced these results uniformly (MCC's between 0.600 and 0.704) practically regardless of the hyperparameters chosen (see also Appendix F)

Catastrophic Forgetting

To preclude the risk of catastrophic forgetting (e.g. figure G2), a warm-up phase of 10% of the training-run was in general included (with both AdamW and Adafactor optimizers). In fact, tests with 0 warm-up were made without any adverse effects, but to avoid time-wasting, the more conservative approach (with warm-up) was generally chosen. It is also recommended to monitor loss during training (e.g. with W&B or Tensor Board) to stop non-productive training-runs.

Figure G2

Catastrophic forgetting with Electra-Large (ema smoothing 50%)