Multi-channel Convolutional Neural Network for Precise Meme Classification

Victoria Sherratt*†
University of Hull
Hull, United Kingdom
v.sherratt-2020@hull.ac.uk

Kevin Pimbblet*‡
University of Hull
Hull, United Kingdom
k.pimbblet@hull.ac.uk

Nina Dethlefs*†
University of Hull
Hull, United Kingdom
n.dethlefs@hull.ac.uk

ABSTRACT

This paper proposes a multi-channel convolutional neural network (MC-CNN) for classifying memes and non-memes. Our architecture is trained and validated on a challenging dataset that includes non-meme formats with textual attributes, which are also circulated online but rarely accounted for in meme classification tasks. Alongside a transfer learning base, two additional channels capture low-level and fundamental features of memes that make them unique from other images with text. We contribute an approach which outperforms previous meme classifiers specifically in live data evaluation, and one that is better able to generalise 'in the wild'. Our research aims to improve accurate collation of meme content to support continued research in meme content analysis, and meme-related sub-tasks such as harmful content detection.

CCS CONCEPTS

• **Information systems** → **Social networks**; • **Computing methodologies** → **Natural language processing**; **Object recognition**; *Transfer learning*.

KEYWORDS

multimodal learning, computer vision and language, neural networks, social media analysis

ACM Reference Format:
Victoria Sherratt, Kevin Pimbblet, and Nina Dethlefs. 2023. Multi-channel Convolutional Neural Network for Precise Meme Classification. In *International Conference on Multimedia Retrieval (ICMR '23), June 12–15, 2023, Thessaloniki, Greece.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3591106.3592275

1 INTRODUCTION

Internet memes are multi-modal content commonly shared online which reference cultural materials, catchphrases, jokes or images to communicate ideas. They exploit external knowledge in combination with text and image modalities to convey meaning; their unique properties make them easily editable or shareable, but difficult to collect or analyse using automated methods.

Internet memes are the focus of ongoing research due to how quickly they circulate online and the complexity of detecting harmful, offensive, hateful or toxic messages in multi-modal content [5, 25, 32]. As meaning is generated through interactions in both modalities, humour and reference to external knowledge, meaning is no longer face value and difficult to decode without context. Varied and comprehensive meme datasets are therefore crucial for such automated content detection tasks.

Currently, available meme datasets are created through extensive manual annotation of collected content to determine whether an image is or is not a meme [29]. In some cases, artificially generated datasets are created for hate-speech detection using the popular superimposed text-over-image meme (image macro) format, which do not represent typical memes shared online that are more varied and contain noisy text [14, 15].

Alternative strategies include collating content from Twitter with the hashtag 'meme', though this approach assumes tagging and categorisation accurately represents that all content is a meme. Additionally, these datasets are 'static' and manual annotation requires updating – in the peculiar case of memes which rapidly evolve and develop new formats, static datasets do not capture emerging memes and may quickly become outdated.

Datasets typically distinguish memes from images such as photographs and do not include other image-with-text (IWT) formats like advertisements, movie posters, online news articles or screenshots of posts which are also circulated online. Thus, models trained on such data perform poorly on live detection tasks, or the subsequent analysis of meme features are not accurate representations of *only* memes and real memes.

1.1 Contributions

We contribute a classifier to distinguish memes from non-memes that is constructed from identified meme features in both modalities. We achieve this with two additional channels alongside a transfer learning base, which are the first to utilise fundamental meme features, namely identifying blank space dominance or low quality editing, text-features for high readability or short text length to conform to restrictive space. These features are indicative of content like memes which are deliberately designed to promote online sharing and user participation through re-editing.

Our proposed approach comprehensively identifies memes and is precise in excluding non-meme IWT formats, particularly in live evaluation tests where previous classifiers have under-performed, demonstrating better generalisation ability than previous work.

*Centre of Excellence for Data Science, AI, and Modelling (DAIM), University of Hull.
†School of Computer Science; Big Data Analytics Research Group, University of Hull.
‡E.A. Milne Centre for Astrophysics, University of Hull.

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06.
https://doi.org/10.1145/3591106.3592275

The data considered for the development of our classifier addresses the research gap of IWT non-memes rarely considered in previous classifiers. Although memes are varied in their format, they also comprise enough common characteristics to be considered collections of content. Our research examines a method of identifying memes more broadly, without focusing on one particular subset of memes or a single meme type, such as image macros which are defined as captioned images (usually bold, impact-font top and bottom text) with typical images used as templates [10].

Amongst studies that do include IWT non-memes, our approach works on images with and without text and does not require external knowledge outside of an image (e.g., captions or user data from Twitter). Further, our histogram channel uses counts of dominant colour palettes, blank space values in an image and overall pixel counts as opposed to histograms or orientated gradients, Haar Wavelet transformations or local binary patterns [24, 30].

We further contribute an analysis of meme features that make them distinct from other non-meme, IWT content, and which is used as the basis for our multi-modal learning architecture. With a more accurate method of classifying memes, we hope to reduce the annotation burden required to acquire large meme datasets for future research. Finally, we propose a classification architecture which in future can be easily adapted to account for new, emerging meme formats and an approach that can classify images with or without text, whether meme or non-meme.

2 RELATED WORK

Meme definitions. We adopt the definition of memes from Shifman's *Memes in the Digital World*, which describes memes as, "a group of digital items sharing common characteristics of content, form, and stance; that [are] created in awareness of one another; that are circulated, imitated, and/or transformed via the internet by users." [34]. As user-generated content, memes are considered products of 'participation' by multiple users, which distinguishes them from other images and IWT combinations; as Shifman [33] further notes, they are "marked as textually incomplete or flawed, thus distinct from and perhaps defiant of glossy corporate content."

In addition, Knobel and Lankshear [16] note that common features of memes are intentionally used to encourage participation via resharing *and* editing, thus meme formats evolve over time from continuous user participation. These attributes make memes distinct from other media content, such as viral videos – which are commonly reshared, but are not edited by participating users. Rogers and Giorgi [28] similarly argues that memes are collections of technical content by analysing memes created using image generators.

There is significant work identifying harmful or offensive memes as part of ongoing research in the detection and prevention of toxic/hateful online content. We therefore split meme classification into two categories: classification tasks concerned with identifying subsets of meme content (e.g., harmful vs non-harmful, propaganda vs truth) and classification concerned with distinguishing memes from non-meme content.

Hate speech detection. Afridi et al. [1] conducted a comprehensive study of multi-modal meme classification approaches, covering both meme vs non-meme classification and other classification tasks. Their survey noted that state-of-the-art multi-modal transformers perform poorly in meme related tasks; the authors suggest that, in standard vision and language tasks like image captioning, efforts are made to generate the best explanation for an image, but there is little semantic alignment between image and text in memes.

Sharma et al. [32] conducted an extensive survey of harmful meme classification and available datasets, noting that the majority of state-of-the-art harmful content classification approaches use similarly large-scale pre-trained neural networks for visual and text content. However, the authors also outlined the complexity of the task and challenges including subjective label annotation, insufficient dataset size and rapid evolution of memes. Whilst our research does not address harmful content, it does aim to improve the availability of meme datasets, reduce annotation burden for meme detection and maintain better accuracy in live evaluation.

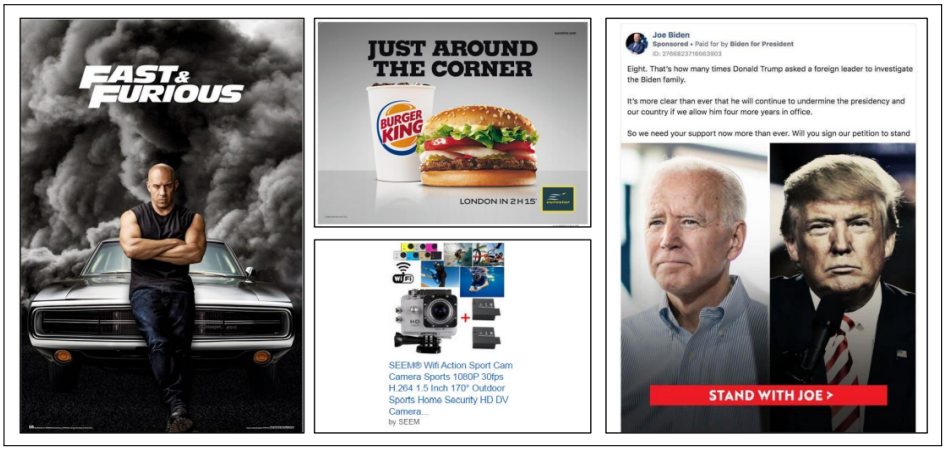
Facebook set a prominent multi-modal classification task to detect harmful memes, 'The Hateful Meme Challenge', with an artificially generated dataset with benign-confounder images to encourage participants to consider both modalities in their solution [14]. Kirk et al. [15] examined the generalisability of these models and noted poor performance on 'wild' memes, mostly due to issues with optical character recognition text extraction as the Hateful Memes competition dataset included generated text as an attribute.

Other classification tasks have included the identification of 'troll' memes [7, 20, 37] with Pramanick et al. [25] introducing finer categories for identifying propaganda techniques. Mookdarsanit [21] proposed an approach to classifying hate speech in non-English memes and Barnes et al. [3] classified popular memes on Reddit.com during the COVID-19 pandemic with a content-analysis approach to identify what features made memes popular.

In most hate speech or sub-category meme detection, IWT non-memes are rarely considered and datasets are either a small sample size or artificially generated, thereby poorly comparing to the variety of memes and non-memes shared online. In contrast, we provide a better representation of memes and non-memes circulated online with our training dataset and do not use multi-modal transformers due to the limitations identified by Kirk et al. [15] and Afridi et al. [1] when employed in meme-related tasks .

Meme vs non-meme detection. Tasks that specifically deal with meme vs non-meme classification tend to employ a variety of approaches outside of the more commonly seen multi-modal transformers in meme 'content' classification. The 2020 EVALITA competition 'DANKMEMES' included a sub-task for meme vs non-meme classification using an annotated dataset with information such as visual actors or image manipulation [19]. However, a drawback of this competition is the small sample size of the dataset and localisation to one particular event. Leskovec et al. [18] presented an earlier 'Meme-tracker' to identify memes via short distinctive phrases through topic modelling and phrase graphics. Whilst the authors did not analyse the visual content of memes, Leskovec et al. [18] demonstrate an approach which is able to distinguish memes via their unique linguistic characteristics.

MemeHunter is a notable meme vs non-meme classifier proposed by Beskow et al. [4], a multi-modal architecture utilising optical character recognition for text extraction, object detection and image similarity. However, their model performed poorly when evaluated against data from the US midterm elections due to more



(a) Non-Meme IWT Formats: movie poster (left), adverts (middle), social media (right)

(b) Example meme formats

Figure 1: Example non-memes IWT formats and typical meme formats.

sophisticated memes online than in their training data. The authors draw particular influence from an earlier meme classifier by Dubey et al. [9], based on meme template matching.

We identify four approaches that include IWT non-memes are part of their training dataset. Du et al. [8] proposed a similar model as MemeHunter, extracting visual features with ResNet50 and the element-wise average GloVe word embeddings from OCR extracted text; the authors also tested their algorithm against the early template matching method proposed by Dubey et al. [9] and found their method outperformed this earlier classifier. Perez-Martin et al. [24] compared meme classification approaches using histogram of orientated gradients, support vector machines and deep learning models with input from visual and textual modalities. Transformer models have also been presented for general meme classification tasks with the inclusion of IWT data [17]. Our approach differs from these by employing three channels separately trained on identified meme features instead of two dedicated streams for image and text only, and a histogram channel not based on orientated gradients but instead colour values, blank space and pixel counts.

Sharma and Pulabaigari [30] present a three-channel approach to meme classification comprising visible feature extraction with VGG16, text feature extraction with emotion detection, and a semantic similarity measure between both modalities. Sharma et al. [31] also contribute a model to distinguish memes using canonical correlation analysis between the text and image modalities of memes and non-memes. In both previously listed studies, the authors convert their dataset to entirely images-with-text by using captions of images when text is absent. In comparison, our approach does not require an image to have text or information external to the image (e.g., captions).

3 DEFINING MEME FEATURES

As indicated in Figure 1, there are numerous types of online content that share meme features but are not memes under Shifman's definition [34]. Non-meme types outlined here are not created, edited or transformed by internet users. They do not belong to a specific group of content (e.g., a subset of memes, such as image macros). Importantly, some of the content is designed to advertise or persuade; they are not opinions of users who created them, but rather the stance of brands intended to sell a product or idea.

As noted by Kirk et al. [15] and in prior research, model performance is less accurate outside of competition or training scenarios due to the variety of IWT formats in online spaces compared to training data. The difficulty in collecting memes relates to their boundaries which are blurred, as memes take materials from existing artefacts and mimic them in form, structure, style, language and design - but repurposed to communicate a different message.

The re-use of some images or catchphrases can be manipulated in ways that, in the context of a meme, carry an entirely different meaning to their origin. When performing deeper analysis of linguistic and visual meme features on datasets that incorrectly contain IWT formats, the subsequent analysis is likely to be a less accurate representation of memes circulated online. Whilst reusing cultural materials and effectively 'mimicking' other online content makes memes harder to detect, it is also this re-purposing and deliberate design to be re-shared and re-edited by other users that provide subtle visual and textual markers used in our architecture to detect memes.

3.1 Data

A baseline model was trained on memes from the Memotion competition [29], memes collected from Reddit.com [3] and non-memes from the Flickr8k data-set [11], with 8,000 images each for memes and non-memes. We trained a convolutional neural network (CNN) on this dataset as a baseline model to compare the impact of excluding IWT images on classifier accuracy. An additional data-set was created to include IWT non-meme formats commonly circulated online; as expected, the baseline classifier achieved significantly poorer performance on this dataset (see Section 5 - Results). The analyses presented in this paper are the results of architectures trained this extended data of of memes and non-memes with the inclusion of IWT formats.

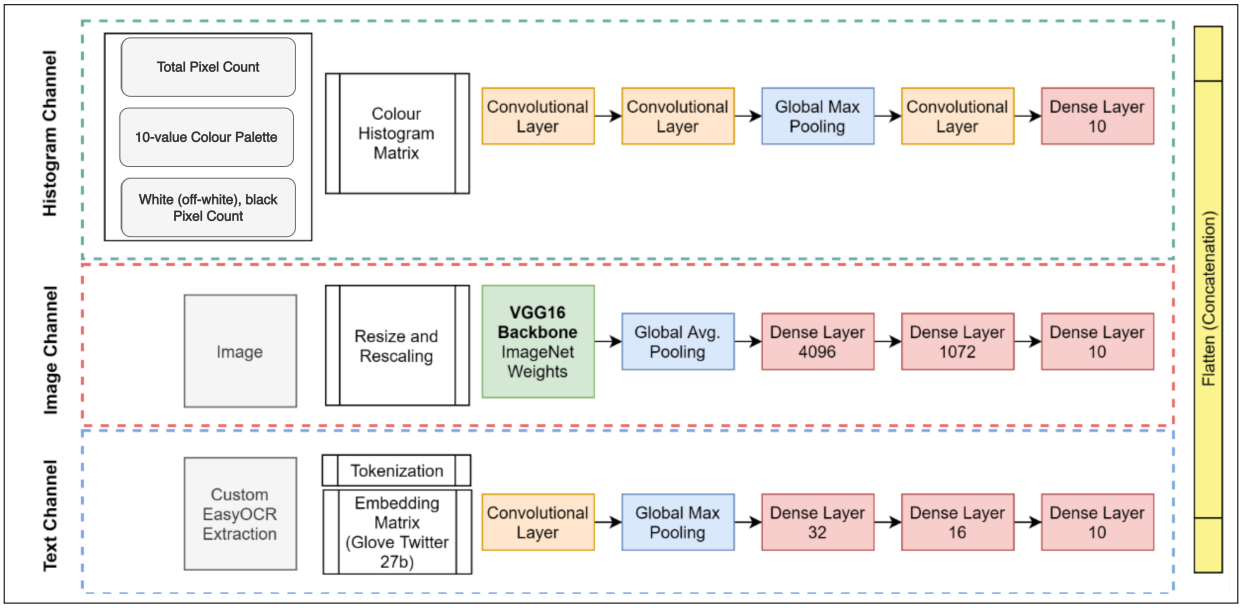


Figure 2: Proposed multi-channel convolutional neural network (MC-CNN) architecture.

Table 1: Data sources for meme classification and live validation. *ADS-16 is a dataset of online adverts.

Data Source	Type	N. Samples
Memotion [29]	Memes	6,992
Memegenerator [2]	Memes	6,718
Reddit [3]	Memes	6,290
Flickr30k [38]	Non-Memes	10,000
ADS-16* [27]	Non-Memes	286
Advertisements [12]	Non-Memes	7,720
IMDB Posters	Non-Memes	1,994
Total	-	40,000
Twitter	Memes	1,367
Twitter	Non-Memes	1,367
Total	-	2,734

The training datasets contain a balanced class of memes and non-memes, extending the original datasets to include IWT non-memes and re-balance available memes from other sources. We use the Flickr30k dataset going forward to increase the size of the original Flickr images by 2,000 images [38]. An additional 10% is kept from all sets for validation outside of training and to compare multiple models. A further set of evaluation data was collated from Twitter to examine model performance on live data.

All data is available from the sources outlined in Table 1. The images for the the IMDB dataset and URLs from the Twitter evaluation dataset are made available in the supplemental materials. Images from the Twitter data in particular are not shared to ensure users retain the right to be forgotten/remove their content from public forums. For the live data evaluation, a balanced set of memes and non-memes were created by collecting Twitter images through the Tweepy package [26].

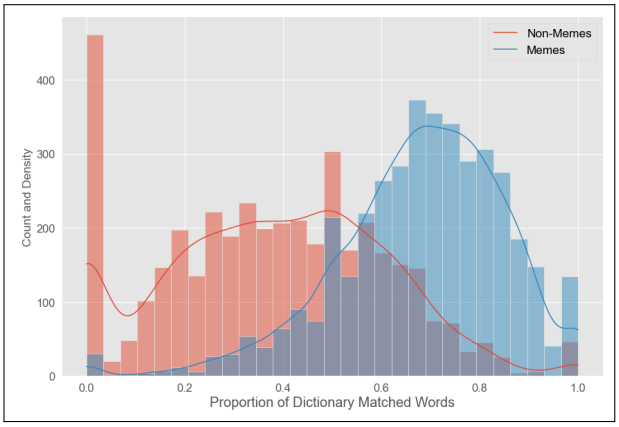
One author of this paper manually annotated these images to indicate which class a sample belonged to; two further annotators labelled 10% of the collated data to measure inter-annotator agreement, where the sample would be increased above 10% if there was found to be little agreement among annotators. We use Cohen's Kappa coefficient [6] to calculate an inter-annotator reliability score of 0.76, indicating three annotators were in substantial agreement.

4 MULTI-CHANNEL APPROACH

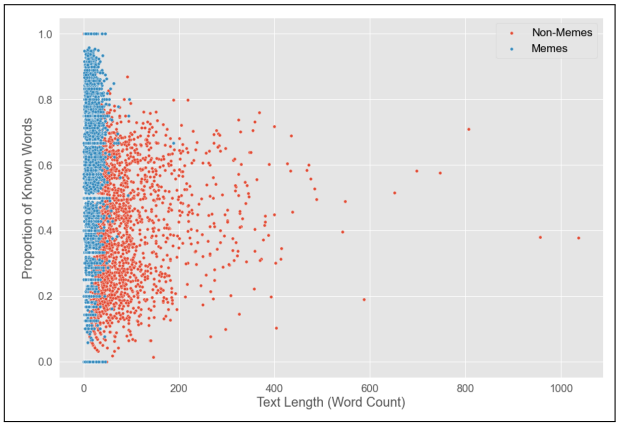
Visual salience analysis of the baseline classifier indicated this model focused on the presence of text in images to classify memes, which would be unlikely to distinguish memes from IWT non-memes. We therefore explored individual text and image channels in more detail to understand which features were important in each modality and could be used alongside the visual features extractor from an image-only CNN.

Given their usual format as one incorporating text, a model based on text-only features examines what textual attributes were unique to memes in comparison to non-meme IWT formats. We also tested variations of histogram channels, including local binary patterns (LBP), histogram of orientated gradients (HOG) and Haar wavelet transformations, which have previously seen promising results [24, 30]. However, these were not used in the final architecture proposed as our histogram variation outperformed these approaches.

Other potential channels were explored, including template matching, however this was deemed less useful as meme formats change over time. Object detection and facial recognition/detection were also ruled out, as in the case of movie posters and adverts individuals who appear in memes may also appear in non-meme



(a) Proportion of extracted text matched in dictionary.



(b) Text length (word count) of memes and non-memes.

Figure 3: Proportion of matched dictionary words and text length (word count).

IWT formats as memes tend to re-appropriate available icons and material.

Instead, we focus on identifying those fundamental meme features that make them different from other meme content and would not become outdated. Whilst "made in awareness of one another" is part of the meme definition described by Shifman [34], this would only be possible with object detection and template matching which is not used for the above reasons. The proposed architecture incorporates three separate channels:

(1) **Text Channel:** Images are passed through the fine-tuned EasyOCR [13] model and text is extracted where available. Tokenized sequences are used as the channel input with a corresponding embedding matrix.
(2) **Image Channel:** VGG16-backbone CNN [35] with ImageNet weights fine-tuned on the meme and non-memes dataset.
(3) **Histogram Channel:** Colour palette values, dominant colour values, total pixel count of white, off-white and black colour values are extracted from images as input for a CNN.

The feature layer of each channel is flattened into a fully connected layer and concatenated. A 2-stage head comprised of two fully connected layers of 100 units with a dropout of 25% between each layer is used before a final binary classification layer. All model architectures are trained on the aforementioned balanced set of memes and non-memes with 70/30 training and test split for 100 epochs with early-stopping, a high-learning rate (0.001) Adam optimizer with learning rate reduced on plateau for after 2 epochs of no improvement to validation loss. As the majority of meme classification tasks leverage transfer learning, we focus on our unique contributions in the text and histogram channels [1, 4, 8, 24, 30, 31].

4.1 Text Channel

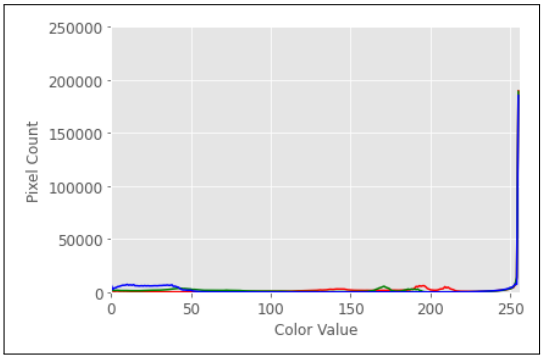
Text is extracted from both memes and non-memes using optical character recognition (OCR). EasyOCR, Keras-OCR and PyTesseract [13, 22, 36] were tested for their accuracy on Memotion and Memegenerator datasets, which include ground-truth for the text in memes. EasyOCR proved the most accurate for extracting text from memes and adverts. To improve the accuracy of meme extracted data, a fine-tuned EasyOCR model was developed using the ground-truth in Memotion and MemeGenerator to generate single words or long sentences with a variety of stroke widths, text fonts and backgrounds.

This fine-tuning corrected common OCR mistakes, namely related to difficulty distinguishing spaces in meme text which used thick stroke widths or Impact font. The details of generating the specific meme-font dataset for retraining OCR models is available in the supplemental materials.

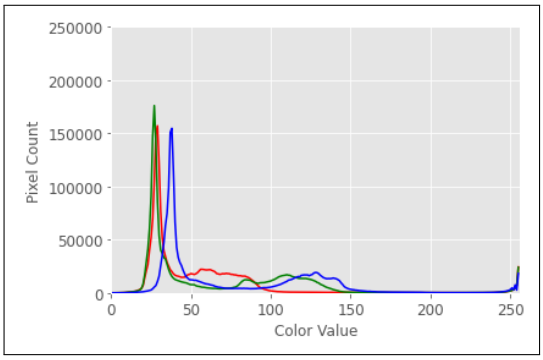
The modified OCR package identifies typical font typefaces found in memes - many of these fonts are part of popular online meme generator websites, where user can create memes without technical knowledge or skill; in other cases, fonts are typical of particular meme types and formats. It was not possible to fine-tune an OCR model for the advertisements and movie posters in the dataset, as there is no available ground-truth for these images.

Still, without a fine-tuned OCR model, extracting text was more accurate and readable with memes than it was with non-meme data. This could be due to the 'low' barrier to participation for memes, which is necessary for their continual dissemination and evolution [34]. Placing text in blank spaces, or superimposing text on an image, reduces the editing skill required to participate in sharing memes compared to the editing skill required to create a movie poster or advertisement; thus, text is both easier to extract and alter in memes, whereas other content is often more complex in design for different purposes.

We use the readability of text in memes as a heuristic for 'high quality' content (text the fine-tuned OCR model struggles to extract) and 'low quality' content (text the OCR has been fine-tuned to extract). Figure 3 (left) shows the proportion of extracted words which are matched to dictionary English words per content type. Extracted meme text produces more decipherable words and non-memes extracted text tends to produce character combinations which do not form recognisable words and therefore are not matched to a dictionary word. Additionally, it is still possible to estimate word lengths; Figure 3 (right) demonstrates memes have overall shorter word



(a) Meme histogram values.



(b) Non-meme histogram values.

Figure 4: Colour channel (red, blue, or green line) pixel count of meme and non-meme samples.

length than adverts and movie posters, though there are exceptions (e.g., memes without text; images without text).

The final architecture for the text channel is a CNN with GloVe word-embeddings as the channel input [23]. We use the pre-trained Twitter GloVe word vectors, as a non-contextual word embedding model since much of the text in our sample is comprised of short sequence lengths and the pre-training domain matches our research area (social media). We use a vocabulary size of 15,000 and padded sequence lengths to a maximum length of 2,000, masked in the embedding step to represent text features and feature length. Images without text were given zero values throughout the embedding matrix.

4.2 Histogram Channel

Other image attributes alongside the features extracted via our transfer-learning image channel were investigated further due to the success of previous research in classifying memes on template formats, or histogram values [9, 24, 30]. However, as Sharma and Pulabaigari [30] note, histogram values alone are not adequate to classify memes. We explore a different approach to determine whether lower level colour features can improve predictions by focusing on the dominance of blank space, colour profiles and pixel counts which are indicative of low-quality editing in memes.

As noted in the previous section, memes are easy to alter to encourage participation, and one strategy to facilitate this is the presence of blank spaces to allow users to add their own text. Considering the memes in Figure 1, this approach has advantages over other feature extraction methods, namely object and facial recognition, where cultural icons are reused in memes.

Typical CNNs are utilised for unsupervised feature extraction and demonstrate excellent performance for object detection; in this case, the goal is to identify the *absence* of features. Further, blank space is not typically restricted to one single region in all memes. The histogram channel inputs are a set of curated colour values in an image that quantify the amount of editable space in an image as a complimentary channel to the image channel. The colour features of images showed significant performance increase both as a single channel compared to the baseline and in combination with other channels.

Figure 4 shows the analysis of colour features in memes and non-memes, where memes typically have a much higher count of white-space than most adverts; they typically have a much lower pixel count overall (again due to their small size and lower quality production to advertisements) and less complex colour palettes compared to advertisements, although this last attribute varies as memes tend to draw on many images or referential material available.

Images are first passed through a function to extract the relevant values per red, blue or green channel to input into a CNN. The numerical input for the histogram channel is the count of pixels in each red, blue and green (RGB) channel for 10 dominant colours (30 values); count of pixels in blank spaces defined as white, off-white, black and off-black shades and finally overall pixel count in each RGB channel.

5 RESULTS

Combinations of individual and dual channels were tested as well as the proposed MC-CNN. The baseline model is also provided for comparison, which is a single-channel CNN without VGG16 as a backbone and trained on a dataset of memes and non-memes without IWT samples.

In Table 2, we compare the accuracy of trained models on 10% of data withheld from the training datasets. In Table 3, we report the the percentage of images each model classified per class against the ground truth label, as well as the calculated F Measure (F1 Score) on the same validation dataset. We also compare our model to the one

Table 2: Model performance comparison. *Baseline model trained on the original datasets containing memes and Flickr8k only.

Model	Training Acc.	Validation Acc.
Baseline*	97.76	53.10
Image-Only	94.28	88.08
Text-Only	94.14	69.38
Histogram-Only	81.04	78.93
Image + Text	94.54	82.08
Image + Histogram	94.55	86.30
Text + Histogram	94.08	64.48
MC-CNN	*97.85*	*92.43*

proposed by Du et al. [8], one of the few trained on IWT images and made available to other researchers at the time of this paper. We report the original published scores and retrain their model on our own data, to understand whether any performance gains are due to architecture or the compilation of data.

Although Sharma and Pulabaigari [30] present a three-channel classifier, the authors superimpose the captions of non-text images to artificially convert all images to IWT formats, whereas captions are not available as a feature in the datasets used in our study. Semantic similarity is calculated from either superimposed caption text or extracted meme text, against a generated scene descriptor of the image; in the case of Flickr8k images, the captions of Flickr8K used for superimposition are descriptions of the image and thus will have a high semantic similarity to a scene descriptor. The authors also train their classifier on a much smaller dataset and is unavailable to re-train with our dataset for a fair comparison.

Table 3: Classification accuracy and F1 on validation dataset.

Model	Prediction	Label		Validation F1
		Meme	Non-Meme	
Baseline	Meme	99.60	93.40	67.99
	Non-Meme	0.40	6.60	
Image-Only	Meme	76.75	0.60	86.55
	Non-Meme	23.25	99.40	
Text-Only	Meme	68.55	29.80	69.12
	Non-Meme	31.45	70.20	
Histogram-Only	Meme	74.60	16.75	77.97
	Non-Meme	25.40	83.25	
Image + Text	Meme	70.05	5.90	79.62
	Non-Meme	29.95	94.10	
Image + Histogram	Meme	86.60	14.00	86.34
	Non-Meme	13.40	86.00	
Text + Histogram	Meme	66.65	37.70	65.23
	Non-Meme	33.35	62.30	
MC-CNN	Meme	96.90	12.50	92.75
	Non-Meme	3.10	87.96	
Du et al. (Original) [8]	-	-	-	73.00
Du et al. (Trained on our data) [8]	-	-	-	*97.11*

Our results in Table 3 indicate the inclusion of a second channel can reduce performance, for example the image + text architecture achieves less performance than the image-only channel. Likely this is due to noise introduced by the text channel, the weakest single-channel architecture. The image channel improves performance significantly in most combinations, whereas the text and histogram channels (alone and in combination) perform poorly without a transfer-learning image base. Notably, the MC-CNN outperforms other channel combinations and excludes the most non-memes from incorrect classification as memes.

Whilst Du et al. [8] outperform the MC-CNN when trained on our dataset, their model was originally trained on heavily manually annotated data, including post-OCR correction; similarly, the training and validation data used to re-train their model in our comparison contains cleaner text for memes in particular due to the availability of ground-truth text labels in Memotion and Memegenerator [2, 29]. In contrast, the low-level text and histogram features used as input for two channels in our architecture requires less clean text (in the case of the histogram channel, text is not required at all). We further examine the performance these models in the next section, where the text input is expected to be noisy.

5.1 Live Data Evaluation

The evaluation on Twitter data demonstrates the MC-CNN and the Image-Only channel achieve better performance as with training results, though as with other studies there is a notable drop in performance on live data. We compare our results to the model Du et al. [8] proposed, trained on our dataset for a fair comparison. We view live data evaluation as a crucial step, as many previous classifiers have achieved good results in training but poor results in live evaluation [4, 15, 30]. Whilst their model performed better in the prior validation test, performance decreased in live data evaluation as all models listed in Table 4 have.

Table 4: Twitter data evaluation of top 4 performing models and related research.

Model	Acc.	Precision	Recall	F1 Score
MC-CNN	**82.15**	76.56	92.68	**83.85**
Image-Only	80.65	**92.38**	66.69	77.46
Image-Histogram	73.81	72.70	76.01	74.32
Image-Text	70.59	77.70	57.52	66.10
Du et al. [Trained on our data] [8]	74.62	67.38	**93.38**	78.38

We suggest the MC-CNN is better able to generalise than comparison models, and that multimodal approaches based only off input text and image is not sufficient to accurately predicted challenging dataset like images circulated on Twitter. Du et al. [8] focused primarily on detecting memes with text from non-memes IWT with image and text input only; in the case of live detection, memes and non-memes can contain both modalities or only one, and in live evaluation text data is likely to be noisier. The additional histogram channel of our classifier performs the same function regardless of whether both modalities exist, and is better able to identify instances of poor image alteration innate to many memes.

For the MC-CNN, non-memes incorrectly classified as memes tended to be examples of user-generated content (e.g., a digital drawing, screenshot of other viral content, a user generated advert or design) but not necessarily considered a meme. Given the original training data did not contain user-generated images and only corporate content, lower performance on this type of data is expected.

There is some difficulty defining memes themselves. For example, the practice of screenshotting and re-sharing humourous content is popular on Twitter, though not necessarily following the principles of altering or editing to make content a meme; however, such content shares features of memes in their format and linguistic attributes. In the Twitter evaluation dataset, these images were not considered a meme.

6 CONCLUSION

In this paper we proposed a multi-channel convolutional neural network for meme and non-meme classification, which outperforms models trained on a similar dataset of IWT non-memes in live data

evaluation. The individual channels that comprise the MC-CNN were developed from analysis of meme text and colour features in relation to IWT non-memes. Whilst we propose an image channel with transfer learning as other models have, the two additional channels exploit different features of memes than previous studies, focusing on the visual and textual markers that make such content 'textually incomplete and flawed' compared to other IWT content [33].

Our architecture retains better performance in live evaluation tests, a crucial step for classifier in meme-related tasks which often perform poorly outside of training [15]. The boundaries between memes and other content is not always clear; memes mimic and reuse cultural materials from other images, and their formats continually evolve through participation. The architecture presented is better able to generalise those varied formats by focusing on the markers of user-edited content rather than image or object detection.

A classifier that can accurately collate more memes would improve tasks relating hate speech detection, harmful content or propaganda detection by increasing the availability of data representative of real memes and facilitating accurate analysis of features that make multi-modal content like memes offensive. Currently, this is less possible when datasets include incorrect IWT formats, as the strategies used by memes to generate meaning are unique to user-generated meme content.

6.1 Future Work

Areas of future work for meme classification should expand the IWT non-meme sample to include a greater representation of non-meme content from other platforms aside from those listed in the Data section, including user-generated non-meme content. Secondly, improvements can be made to the text-channel of the architecture with better OCR extraction from non-meme text formats in live data, to incorporate higher level textual features than available with current OCR extraction methods. In our experiments, data available from the IWT non-memes could not be interpreted beyond low-level text features due to poor OCR extraction. Whilst this worked as a novel feature for our method, we also note other researchers have previously succeeded in identifying memes from linguistic features alone [18].

We anticipate the proposed architecture can be adapted to relevant specific meme sub-tasks - for example, focusing on text-only memes - offering flexibility to collate content which has evolving boundaries both in agreed upon definition and new content that emerges. Future work would consider how these features are weighted to consider varying meme formats.

We use the definition of previous social science research to establish what would be considered a meme. The practice of screenshotting and sharing content, particularly with additional comments added by a user, may be of consideration for research in online communication and hate speech detection, and could be considered a type of meme-like practice. However, the development of better definitions of meme content and sub-genres of meme-like practices may start with a more precise methods of filtering out IWT non-meme content, as the classifier proposed in this paper works towards.

REFERENCES

[1] Tariq Habib Afridi, Aftab Alam, Muhammad Numan Khan, Jawad Khan, and Young-Koo Lee. 2020. A Multimodal Memes Classification: A Survey and Open Research Issues. (Sept. 2020). https://doi.org/10.48550/arXiv.2009.08395 arXiv:https://arxiv.org/abs/2009.08395

[2] Library of Congress American Folklore Centre. [n. d.]. Meme Generator: collected datasets. Available at: https://www.loc.gov/item/2018655320/ (2022-05-10).

[3] Kate Barnes, Tiernon Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. 2021. Dank or not? Analyzing and predicting the popularity of memes on Reddit. *Applied Network Science* 6, 1 (March 2021). https://doi.org/10.1007/s41109-021-00358-7

[4] David M. Beskow, Sumeet Kumar, and Kathleen M. Carley. 2020. The evolution of political memes: Detecting and characterizing internet memes with multi-modal deep learning. *Information Processing & Management* 57, 2 (March 2020), 102170. https://doi.org/10.1016/j.ipm.2019.102170 Number: 2.

[5] Tanmoy Chakraborty and Sarah Masud. 2022. Nipping in the bud: detection, diffusion and mitigation of hate speech on social media. *ACM SIGWEB Newsletter* Winter (2022), 1–9.

[6] Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement* 20, 1 (1960), 37–46.

[7] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting Propaganda Techniques in Memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Online, 6603–6617. https://doi.org/10.18653/v1/2021.acl-long.516

[8] Yuhao Du, Muhammad Aamir Masood, and Kenneth Joseph. 2020. Understanding Visual Memes: An Empirical Analysis of Text Superimposed on Memes Shared on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 153–164. https://doi.org/10.1609/icwsm.v14i1.7287

[9] Abhimanyu Dubey, Esteban Moro, Manuel Cebrian, and Iyad Rahwan. 2018. MemeSequencer: Sparse Matching for Embedding Image Macros. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. ACM Press, Lyon, France, 1225–1235. https://doi.org/10.1145/3178876.3186021

[10] Marta Dynel. 2016. "I has seen image macros!" Advice animals memes as visual-verbal jokes. *International Journal of Communication* 10 (2016), 29.

[11] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Flickr8k Dataset. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[12] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (Honolulu, HI, USA). IEEE, 1705–1715. https://doi.org/10.1109/CVPR.2017.123

[13] JaidedOCR. 2022. EasyOCR. Available at: https://www.jaided.ai/easyocr/ (2022-05-10).

[14] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. In *Advances in Neural Information Processing Systems NIPS'20* (Vancouver, BC, Canada), Vol. 33. Curran Associates, Inc., Red Hook, NY, USA, Article 220, 2611–2624 pages.

[15] Hannah Kirk, Yennie Jun, Paulius Rauba, Gal Wachtel, Ruining Li, Xingjian Bai, Noah Broestl, Martin Doff-Sotta, Aleksandar Shtedritski, and Yuki M Asano. 2021. Memes in the Wild: Assessing the Generalizability of the Hateful Memes Challenge Dataset. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*. Association for Computational Linguistics, Online, 26–35. https://doi.org/10.18653/v1/2021.woah-1.4

[16] Michele Knobel and Colin Lankshear. 2007. Online memes, affinities, and cultural production. *A new literacies sampler* 29 (2007), 199–227. Publisher: New York.

[17] Christos Koutlis, Manos Schinas, and Symeon Papadopoulos. 2023. MemeTector: Enforcing deep focus for meme detection. *International Journal of Multimedia Information Retrieval* (Jan. 2023). https://doi.org/10.5281/zenodo.7554267

[18] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the dynamics of the news cycle. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 497–506.

[19] Martina Miliani, Giulia Giorgi, Ilir Rama, Guido Anselmi, and Gianluca E. Lebani. 2020. DANKMEMES @ EVALITA 2020: The Memeing of Life: Memes, Multimodality and Politics. In *EVALITA*. http://ceur-ws.org/Vol-2765/paper174.pdf

[20] Ankit Kumar Mishra and Sunil Saumya. 2021. IIIT_DWD@EACL2021: Identifying Troll Meme in Tamil using a hybrid deep learning approach. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv, 243–248. https://aclanthology.org/2021.dravidianlangtech-1.33

[21] Lawankorn Mookdarsanit and Pakpoom Mookdarsanit. 2021. Combating the hate speech in Thai textual memes. *Indonesian Journal of Electrical Engineering and Computer Science* 21, 3 (March 2021), 1493–1502. https://doi.org/10.11591/ijeecs.v21.i3.pp1493-1502 Number: 3.

[22] Fausto Morales. 2019. Keras-OCR. Available at: https://keras-ocr.readthedocs.io/ (2022-02-05).

[23] Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[24] Jesus Perez-Martin, Benjamin Bustos, and Magdalena Saldana. 2020. Semantic Search of Memes on Twitter. (Feb. 2020). https://doi.org/10.48550/arXiv.2002.01462 arXiv:https://arxiv.org/abs/2002.01462v4

[25] Shraman Pramanick, Dimitar Dimitrov, Rituparna Mukherjee, Shivam Sharma, Md. Shad Akhtar, Preslav Nakov, and Tanmoy Chakraborty. 2021. Detecting Harmful Memes and Their Targets. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, Online, 2783–2796. https://doi.org/10.18653/v1/2021.findings-acl.246

[26] Joshua Roesslein. 2020. Tweepy: Twitter for Python! *URL: https://github.com/tweepy/tweepy* (2020).

[27] Giorgio Roffo and Alessandro Vinciarelli. 2016. Personality in computational advertising: A benchmark. In *4th Workshop on Emotions and Personality in Personalized Systems*. 18.

[28] Richard Rogers and Giulia Giorgi. 2023. What is a meme, technically speaking? *Information, Communication & Society* 0, 0 (2023), 1–19. https://doi.org/10.1080/1369118X.2023.2174790 arXiv:https://doi.org/10.1080/1369118X.2023.2174790

[29] Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*. International Committee for Computational Linguistics, Barcelona (online), 759–773. https://doi.org/10.18653/v1/2020.semeval-1.99

[30] Chhavi Sharma and Viswanath Pulabaigari. 2020. A Curious Case of Meme Detection: An Investigative Study. In *Proceedings of the 16th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and Technology Publications, Budapest, Hungary, 327–338. https://doi.org/10.5220/0010110203270338

[31] Chhavi Sharma, Viswanath Pulabaigari, and Amitava Das. 2020. Meme vs. Non-meme Classification using Visuo-linguistic Association:. In *Proceedings of the 16th International Conference on Web Information Systems and Technologies*. SCITEPRESS - Science and Technology Publications, Budapest, Hungary, 353–360. https://doi.org/10.5220/0010176303530360

[32] Shivam Sharma, Firoj Alam, Md. Shad Akhtar, Dimitar Dimitrov, Giovanni Da San Martino, Hamed Firooz, Alon Halevy, Fabrizio Silvestri, Preslav Nakov, and Tanmoy Chakraborty. 2022. Detecting and Understanding Harmful Memes: A Survey. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. International Joint Conferences on Artificial Intelligence Organization, 5597–5606. https://doi.org/10.24963/ijcai.2022/781 Survey Track.

[33] Limor Shifman. 2012. An anatomy of a YouTube meme. *New media & society* 14, 2 (2012), 187–203. Publisher: Sage Publications Sage UK: London, England.

[34] Limor Shifman. 2013. Memes in a digital world: Reconciling with a conceptual troublemaker. *Journal of computer-mediated communication* 18, 3 (2013), 362–377. Publisher: Oxford University Press Oxford, UK.

[35] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *CoRR* abs/1409.1556 (2014). http://arxiv.org/abs/1409.1556

[36] Ray Smith. 2007. An overview of the Tesseract OCR engine. In *Ninth international conference on document analysis and recognition (ICDAR 2007)*, Vol. 2. IEEE, 629–633.

[37] Shardul Suryawanshi and Bharathi Raja Chakravarthi. 2021. Findings of the Shared Task on Troll Meme Classification in Tamil. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*. Association for Computational Linguistics, Kyiv, 126–132. https://aclanthology.org/2021.dravidianlangtech-1.16

[38] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.