Emotion Classification in Newspaper Headlines for Human Behaviour Prediction

Anonymous ACL submission

Abstract

With the advent of Large Language Models (LLMs), substantial performance improvements have been reported across various Natural Language Processing (NLP) domains. However, further evaluation within specific NLP subdomains remains necessary. This study investigates the effectiveness of GPT-based ada-002 text embeddings in conjunction with lexical features for emotion classification in newspaper headlines. Newspapers are chosen, as prior research on emotion detection has largely concentrated on identifying the emotions expressed by the author of a text. Instead, the present study shifts the focus towards detecting emotional responses of the message recipient. To facilitate cross-linguistic comparability, an automated translation procedure is proposed and implemented on three publicly available, labeled emotion datasets to train supervised emotion classification models in both English and German. For both languages, the trained classifiers significantly outperform previous benchmark results by over 42%, demonstrating the superiority of the chosen approach. In terms of language comparison, the English classifier achieves a higher performance with an F1 score of 0.683 compared to an F1 score of 0.655 of the German classifier. To demonstrate the impact of emotional appeal on human behaviour in online marketing, the German classifier is applied to a real advertisement setting. This application reveals how emotional priming detected in the newspaper headline influences the likelihood of user interaction with the advertisements placed within the newspaper article.

1 Introduction

Recent advancements in transformer-based Large Language Models (LLM), notably exemplified by OpenAI's GPT3 (Brown et al., 2020), GPT4 (OpenAI et al., 2024) and Google's Gemini (Team et al., 2023), have surpassed previous state-of-the-art methodologies in most Natural Language Processing (NLP) tasks (Minaee et al., 2024). Given the need to reassess previous approaches and findings in light of these advancements, this study focuses on classifying readers' emotional responses to written text. The digitized marketplace represents a critical area of interest for leveraging emotional appeal in textual content, as emotional engagement plays a pivotal role in influencing human behaviour and marketing effectiveness (Choi et al., 2016; Poels and Dewitte, 2019). Therefore, the domain of online marketing is chosen to showcase the practical relevance of detecting reader emotion.

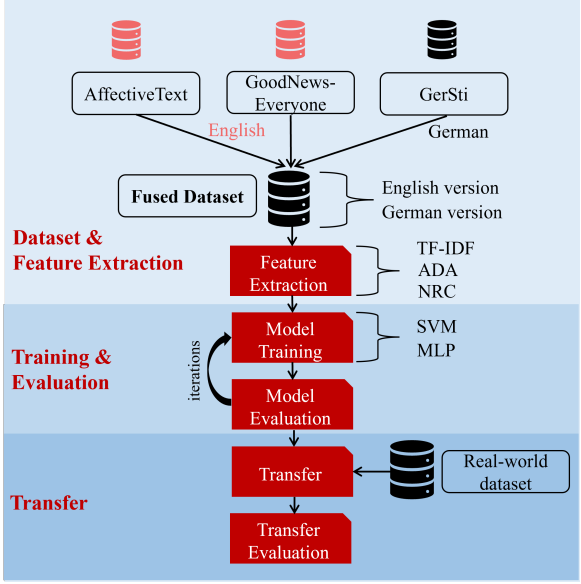


Figure 1: Outline of the study method, separated into a dataset, feature extraction, training, evaluation and transfer phase

Whereas traditional advertisement revolves around addressing wide ranges of potential customers by trying to appeal to a cross-section of society, the digitized marketplace allows for a highly individualized targeting, aiming for an optimal advertisement-person fit (Strauss et al., 2005). Consequently, online marketing is faced with the core challenge of identifying and addressing user pref-

erences to tailor advertisements accordingly. Since user preference itself is a latent construct that is hard to measure, it is commonly approximated by human behaviour (Otamendi and Martín, 2020).

In this work, emotion is considered an anticipatory factor in consumer action tendencies due to its assumed link to human decision making and behaviour, as supported by psychological and evolutionary theories. In psychology, the *Functional Theory of Emotions* states that situational emotional appraisals lead to certain states of action readiness, enabling adaptive behavioural responses (Frijda, 1986). Evolutionary psychology postulates that different emotions emerged as responses to specific adaptive problems relevant in the ancestral environment, leading to distinct action tendencies depending on the experienced emotion (Cosmides and Tooby, 2000). Therefore, this work aims to classify emotional appeal in newspaper headlines in order to anticipate consumer behaviour.

This paper (a) combines and translates preexisting labeled datasets for emotion detection tailored to online marketing, (b) provides benchmark results for the classification of reader emotion by training supervised models on LLM-derived text embeddings and lexical emotion features, (c) examines the applicability and generalization of the resulting model to other languages by comparing classification results for German and English texts and (d) analyzes the expected emotional response pattern in online marketing by applying the trained classifier to a German advertisement use case where data on user activation is available.

2 Related Work

2.1 Advancements in Text Emotion Classification

Research interest in emotion classification based on text has been prominent in the field of NLP for the past decades (Canales and Martínez-Barco, 2014). Early work was dominated by lexicon-based approaches (Hartmann et al., 2019) that primarily involved counting the number of words relating to a certain emotion. The overall emotion of a text is then determined by identifying the emotion with the maximum emotion word count. The NRC Word-Emotion Association lexicon[1] is the largest publicly available emotion lexicon to date, covering over 14,000 English words assigned to one or more of eight possible emotions. Due to its size and availability in 108 languages, it is a popular choice among NLP researchers for emotion classification tasks. Although lexicon-based approaches are a straightforward and rule-based algorithm, they are constrained by their lack of contextual sensitivity and are dependent on the quality and comprehensiveness of their underlying lexicon. This limitation often results in challenges when addressing language-specific jargon (Bandhakavi et al., 2016).

In addition to lexical approaches, the usage of text embedding techniques has come to dominate the field of text classification and continues to be steadily refined. Early approaches predominantly rely on Bag-of-Words (BoW) techniques, which disregard the sequential order of words in a text and use a word-to-document matrix that records frequency counts, often weighted by the word's overall frequency within the text corpus. The most notable example of BoW techniques is the Term Frequency–Inverse Document Frequency (TF-IDF) algorithm (Jones, 1972). While TF-IDF captures more information at the word level than lexicon-based approaches by assigning a static weight to each word's importance across the corpus, it still cannot incorporate contextual information, as it fails to define semantic links between words. To address this limitation, Facebook's FastText model (Bojanowski et al., 2017) was an early adopter of deep neural network architectures to use n-gram character sequences rather than as isolated tokens as input. This method enables FastText to capture morphological information about words and handle out-of-vocabulary words more effectively compared to traditional BoW models, providing a lightweight model to capture contextual dependencies.

In a recent evolutionary step, the development of transformer-based architectures (Vaswani et al., 2017) has laid the foundation for the current state-of-the-art performance of LLMs. These architectures enable the processing of input sequences of variable length while effectively capturing long-range dependencies and contextual information, resulting in context-sensitive and semantically rich vector representations. Notable examples include Google's BERT (Devlin, 2018) and OpenAI's ada-002 (ADA) model, which surpasses the previously leading OpenAI embedding model, text-similarity-davinci-001, in most tasks related to text similarity and classification[2]. The enhanced capabilities of

[1] National Research Council, 2011

[2] OpenAI blog

LLMs also facilitate in-context learning in addition to embedding extraction, where models generate predictions based on contexts augmented with only a few examples (Dong et al., 2024). While this new paradigm addresses traditional challenges in machine learning, such as the need for large training datasets, it tends to underperform in tasks like text emotion classification when compared to fine-tuned, supervised models (Basile et al., 2021).

2.2 Recent Approaches to Emotion Classification

In line with the outlined development in the field of NLP, the integration of embeddings into classification approaches has gained prominence in recent years. Gupta and Yang (2018) proposed the *CrystalFeel* model, which predicts the four emotions *fear*, *anger*, *sadness* and *joy*. The team utilized a total of 7,100 manually annotated English tweets for training. The proposed approach relies on a feature combination of FastText embeddings, TF-IDF embeddings, lexicon-based and linguistic[3] features. The resulting model performed best when all feature sets were included, with the lexicon-based feature and FastText embeddings yielding the best predictive performance. By contrast, the TF-IDF embeddings showed lower predictive power, while linguistic features contributed negligible predictive capabilities to the model.

Babanejad et al. (2019) investigated the importance of emotions expressed within 360,000 Canadian newspaper articles for predicting user interest in the read article. In addition to TF-IDF embeddings, the model relies on a set of linguistic and lexicon-based features. The experiments indicate that the inclusion of these emotion features contributes substantially to the model's performance in predicting user interest, suggesting that the chosen features are effective at capturing and representing emotional content based on newspaper articles.

With the advent of transformer-based text processing, Adoma et al. (2020) evaluated BERT's (Devlin, 2018) capabilities for emotion classification on the English ISEAR (Scherer and Wallbott, 1994) dataset, consisting of over 7,000 emotional expressions from the author's perspective. For the task of classifying seven distinct emotions, BERT outperformed previous classification results on this dataset, demonstrating the promising potential of transformer-based embeddings.

[3]linguistic markers like punctuation or orthography

Recently, Kheiri and Karimi (2023) evaluated the performance of GPT-based embeddings for sentiment classification. Building on the ADA model, the team trained a classifier on 60,000 labelled sentiment tweets. The conducted evaluation revealed that the chosen approach (F1 = 0.87) outperformed BERT embeddings (F1 = 0.73), that were utilized as a comparative benchmark, substantially. In their future work section, the authors propose further investigation of ADA embeddings, specifically calling for their application to emotion detection.

In summary, past work on textual emotion detection has demonstrated the potential of feature combinations, especially when combining embeddings with lexicon-derived features (Gupta and Yang, 2018; Babanejad et al., 2019). Among these embeddings, transformer-based embeddings show the best performance, with ADA embeddings outperforming BERT embeddings for sentiment classification (Kheiri and Karimi, 2023). This observation allows for the hypothesis that ADA embeddings exhibit equally superior performance for emotion classification. However, it can be noted that former approaches rely mostly on easily available data, such as twitter posts. Consequently, a significant imbalance is noted with a predominant focus on the author's emotions, while the recipient's perspective receives limited attention. Exceptions, such as the work of Babanejad et al. (2019) utilize newspaper articles and thereby do account for the reader's perspective, but do not provide emotion labels and only leverage emotional features to predict user interest. Thus, this study addresses this research gap by predicting readers' emotional responses to newspaper headlines through the training of a supervised classifier, utilizing a synthesis of ADA embeddings and lexical NRC features.

3 Emotion Classification Framework

This chapter provides details for the developed Emotion Classification Framework as illustrated in Figure 1, consists of a dataset creation, feature extraction, model training & evaluation and transfer phase. The final transfer of the trained emotion emotion classifier to real advertisement data will be described separately in chapter 5.

3.1 Dataset

Emotion taxonomy. For practical and analytical applications, most emotion classification tasks rely on the concept of base emotions and build on vari-

ations of either Ekman's or Plutchik's emotion taxonomies (Ekman, 1992; Plutchik, 1984). A prominent example is the aforementioned *CrystalFeel* (Gupta and Yang, 2018) approach, which limits its consideration to the four emotions *fear, anger, joy* and *sadness* as they are included in both emotion taxonomies and are therefore adopted for this work.

Dataset selection. For this work, only labeled emotion datasets of newspaper headlines were considered, as they (a) focus on the reader's emotional state rather than the author's and (b) represent prototypical content in which advertisements are placed. Focusing on the headline for emotion extraction is commonly recommended in emotion detection research. Gupta and Yang (2018) argue that the purpose of a newspaper is to catch the reader's interest, which can primarily be achieved by concentrating emotional content in the headlines. This is supported by Strapparava and Mihalcea (2007) arguing that headlines are designed to provoke emotions, making them particularly suitable for automated emotion recognition.

With respect to the aforementioned restrictions, three datasets were identified: *AffectiveText2007*, by Strapparava and Mihalcea (2007) consisting of 1,050 English news headlines covering six emotions; *GerSti*, published by Dang et al. (2021) consisting of 1,000 German headlines labeled into nine emotion and a neutral class and *GoodNewsEveryone*, published by Bostan et al. (2019) and consisting of 5,000 English headlines labeled into 15 emotion categories and a neutral class.

Dataset translation. Due to the mixture of English and German data sources, an automated translation strategy that allows for the evaluation of translation quality is required. For scientific applications, it is common practice to employ a back-translation strategy (Sahin and Dungan, 2014). After translating the text into the target language, a back-translation into the original language is performed. Every pair of original and back-translated text is subsequently evaluated manually by an expert, judging on semantic equivalence.

Since manual evaluation is unfeasible for large datasets, a novel automatic evaluation of translation results is proposed, building on the principles of the back-translation approach. As a first step, translation is performed using a python API to access the Bing translation service[4]. The resulting translation is back-translated, and BERT embeddings are computed for both the original and back-translated texts. To replace the expert's judgement, cosine similarities for every pair of original and back-translated embedding vectors are computed and tested against a threshold of 0.85.

Applying this approach, 86,4% of *GerSti* and 94% of *AffectiveText2007* samples met the threshold requirement for the German dataset. For English data, 99% of *GoodNewsEveryone* samples could be included, resulting in final dataset sizes of 3,462 headlines for the English and 3,167 samples for the German dataset.

3.2 Feature Extraction

In terms of preprocessing, tokenization, lemmatization, stop-word and special-character removal were performed. TF-IDF features (Jones, 1972) were extracted from the preprocessed text by computing uni- and bigrams, limited to the 1,000 most frequent occurrences. ADA embeddings were inferred using the OpenAI API to access the text-embedding-ada-002 model which returns a 1,536-dimensional embedding representation of text input. The features were extracted directly from the unprocessed, raw text to align with the data structure on which the ADA model was trained (Kaplan et al., 2020).

To extract lexical features, language specific versions of the NRC Word-Emotion Association Lexicon were used on the preprocessed input text. In total, 30 NRC features were extracted. The lexicon includes eight emotions as well as two classes that capture positive or negative word connotations. To extract features, relative frequencies of emotion words for each of the ten categories were counted, accounting for different headline lengths by normalizing the resulting feature to a solution space of 0-1. Additionally, for every word in a headline and for every word in the NRC lexicon, FastText embeddings were computed, as they allow for good semantic representation at the word level (Bojanowski et al., 2017). The NRC embeddings were used to compute a class centroid embedding for each of the ten classes, encapsulating a vectorized representation for each emotion class. For each headline, the average distance of all input words to the class centroids were computed using cosine similarities, resulting in additional ten features.

To capture information on mixed emotions, four new classes were computed. The first class evaluates whether a word is only assigned to one emotion by the NRC, whereas the second class measures if

[4] translators

Classifier (Features)	English		German	
	Train F1	Val F1	Train F1	Val F1
SVM (TF_IDF)	0.948	0.467	0.923	0.430
SVM (NRC)	0.621	0.413	0.610	0.412
SVM (ADA)	0.999	0.699	0.981	0.587
SVM (TF_IDF&NRC)	*0.945*	*0.480*	*0.935*	*0.451*
SVM (TF_IDF&ADA)	0.987	0.657	0.985	0.600
SVM (ADA&NRC)	0.999	**0.701**	0.980	0.601
SVM (TF_IDF&ADA&NRC)	0.987	0.667	0.985	0.601
MLP (ADA&NRC)	0.832	0.665	0.820	**0.617**

Table 1: Performances for feature and model combinations on German and English datasets. The baseline from Babanejad et al. (2019) is highlighted in italics, and the best performance per language is indicated in bold.

any emotion assignment is present at all. The third and fourth classes evaluate if a word is assigned to at least two or at least four emotion classes simultaneously. For the four new classes, the same feature extraction approaches based on word count and embedding centroids were applied, increasing the feature size to 28. Lastly, the position of the first occurrence of a positively or a negatively labelled word was measured, resulting in a final NRC feature size of 30.

3.3 Model Training

To identify an optimal classifier, feature combinations were tested systematically. For this procedure, Support Vector Machines (SVM) were implemented for the baseline experiments and were subsequently complemented by a multilayer perceptron neural network (MLP) for the best identified feature combination. These supervised models were chosen, as they are commonly used for text emotion classification (Gupta and Yang (2018); Gambino and Calvo (2019); Kheiri and Karimi (2023)). As incremental effects are expected from feature combinations (Babanejad et al. (2019) and Gupta and Yang (2018)), a combination of TF-IDF and lexical features as introduced by Babanejad et al. (2019), was used for benchmark comparison.

Table 1 summarizes the results of the iterative model training and feature selection process for both English and German datasets. The upper part of the Table reveals that TF-IDF and NRC features, when in isolation, exhibit equal performance with no significant differences between languages. In contrast, ADA features, when used in isolation, (a) outperform the afrementioned isolated features considerably ($\Delta F1_{English\_isolated} = 0.232$, $\Delta F1_{German\_isolated} = 0.157$) and (b) demonstrate varying capabilities depending on the language, as reflected by their superior performance on English data ($\Delta F1_{English\_German} = 0.112$).

The middle part of Table 1 reports on classification results for feature combinations. Compared to the use of isolated features, a further improvement in performance is observed. The highest performance is achieved by combining ADA and NRC features. Additionally, training a shallow MLP on the identified optimal feature combination improved performance on the German dataset, but did not yield any additional benefits for the English dataset. Consequently, a SVM with ADA & NRC features is identified as the best classifier for English data (F1 = 0.701) and a MLP with identical feature configuration for German text (F1 = 0.617).

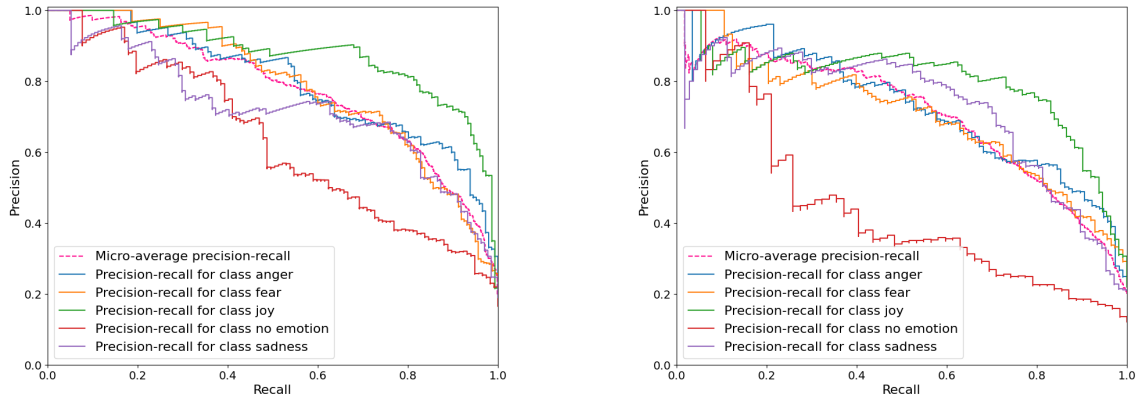
4 Results and Model Evaluation

This section provides final performance metrics of the language-specific model configurations that were identified in section 3.3 on the test dataset.

To examine the classifiers abilities to differentiate between emotion classes, Precision-Recall (PR) curves and F1 scores are presented in Figure 2 and Table 2 at the class level for both language settings.

Emotion	English	German
	Test F1	
No Emotion	0.515	0.479
Joy	0.798	0.758
Anger	0.711	0.667
Sadness	0.664	0.612
Fear	0.721	0.685

Table 2: Test data performance of the respectively best classifier for both languages (SVM (ADA&NRC) for german and MLP (ADA&NRC) for english data) on class level



(a) English classifier

(b) German classifier

Figure 2: Precision-Recall curves on emotion class level for both English and German classifiers.

The PR curves reveal a consistent pattern in both cases, as shown in Figures 2a and 2b: The neutral category *no emotion* proves to be the most challenging to distinguish, as evidenced by the worst-performing PR curves and the lowest F1 scores for both languages. The three emotions *sadness*, *fear*, and *anger* perform comparably well and significantly better than the neutral class, with PR curves closest to the micro-averaged PR curve for the entire classifier. The emotion *joy* stands out as the best-performing class, being the only positively connoted emotion among the base emotions. This distinction in emotional polarity may enhance its predictive potential, allowing the classifier to more accurately assign any positively connoted emotion to the class. In contrast, *anger*, *fear*, and *sadness* are more likely to be closely intertwined due to their primarily negative connotations. The performance of the language-specific models exhibits two notable differences: (a) The English classifier achieves superior performance at the class level, as shown in Table 2. (b) For the German classifier it is harder to detect the *no emotion* class, which is illustrated by its low PR curve in Figure 2b.

Classifier (Features)	**English**	**German**
	Test F1	
SVM (TF_IDF&NRC)	*0.480*	*0.461*
SVM (ADA&NRC)	**0.683**	0.641
MLP (ADA&NRC)	0.678	**0.655**

Table 3: Test data performance for the final model for both languages in comparison to baseline results of Babanejad et al. (2019). Best performances per language are highlighted in bold.

To summarize the overall classification results, Table 3 compares the language specific model performance to the baseline approach of Babanejad et al. (2019). As detailed in Section 3.3, the SVM demonstrates the best performance for the English dataset, while the MLP excels for the German dataset. Consequently, both model architectures were applied to the test dataset. As emphasized in bold, these language-specific model configurations also achieve the best results on the test data. Both models significantly outperform the baseline approach, with an improvement of $\Delta F1 = 0.203$ for the English classifier and $\Delta F1 = 0.194$ for the German classifier, representing a relative performance increase of over 42%. Additionally, the language-dependent effect of increased performance for English text, already found during the training process on validation data, is replicated with a $\Delta F1$ of 0.028.

5 Model Transfer

In the final phase, the resulting German emotion classifier was used to predict emotions for 4,528 newspaper headlines that had been used for advertisement and were published in the second half of 2023. For these headlines, user interactions with the advertisements are recorded. An interaction is considered an *impression* once the advertisement has fully appeared within the reader's visual field, allowing for the actual processing of the advertisement contents. For each recorded impression, user interaction with the advertisement is measured by a click on the advertisement canvas. The average click rate across all sessions is called the Click-

Through-Rate (CTR) and represents the percentage of all instances that resulted in a click on the advertisement. To evaluate the impact of emotional exposure on user behaviour in online marketing, CTR differences between emotion classes are analyzed.

5.1 Quantitative Evaluation

The resulting class distribution reveals considerable imbalance: 74.7% (3,383) of the headlines were classified as *no emotion*, 15.3% (695) as *joy*, 4.6% (206) as *anger*, 2.9% (133) as *sadness* and 2.5% (111) as *fear*. Given the high variance in user interaction rates for each article (mean=146 impressions per article with std=2,336), the total impression counts for each class are illustrated in Figure 3, showing a slight divergence from the article class distribution (*no emotion*: 58%, *joy*: 30.1%, *anger*: 7.3%, *sadness*: 1.9%, *fear*: 2.7%), while maintaining the overall trend of class imbalance.

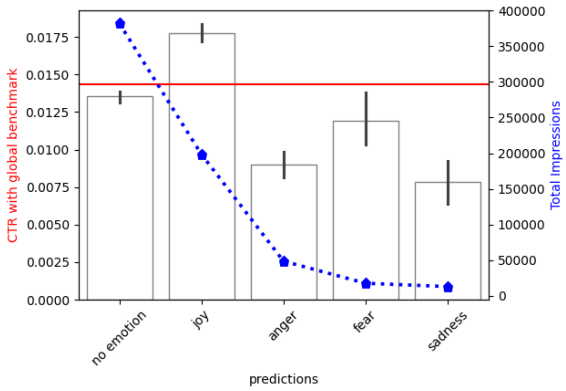


Figure 3: Application of the classifier to German news headlines used for advertisement. Left x-axis illustrates class level CTR, right x-axis sums the total amount of measured, visible impressions. Global benchmark refers to the average CTR over all emotion categories.

Figure 3 illustrates the CTR for each emotion class based on all measured impressions. Notably, the neutral class closely aligns with the average CTR (global benchmark), which is expected due to its predominance in the dataset. Aside from the neutral class, the categorization of emotions introduces notable variance to the CTR, highlighted by two key observations: (a) Best performance is achieved by the positively connoted emotion *joy*, surpassing the average CTR of 1.45% by a Δ of 0.3%. (b) Negatively connoted emotions seem to result in lower interaction rates, with all falling below the average CTR. Among them, only *fear* comes close to the average CTR, whereas *anger* and *sadness* yield interaction rates below 1%.

Besides the consideration of positive or negative emotional tone, the findings also support claims of varying interaction rates in social media depending on the arousal an emotion induces. Prior work by Berger and Milkman (Berger, 2013; Berger and Milkman, 2012) indicate that emotions associated with high arousal (such as *anger*, *fear*, *excitement* or *amusement*) lead to higher interaction likelihoods in internet communication. To apply this to the current study and the used emotion classes, *joy*, *anger* and *fear* can be considered emotions with high arousal potential. Empirically, all of them show higher interaction rates compared to the emotion with low arousal potential (*sadness*).

5.2 Qualitative Evaluation

Since the model transfer involves applying the classifier to unlabeled data, a proper and quantitative evaluation of prediction quality is not possible. Instead, this section presents illustrative cases to highlight specific classification challenges, segregated into three identified areas.

Introspection and Protagonist-Reader confusion. Identifying the emotion an article evokes in the reader poses inherent challenges. The classifier must distinguish between the reader's emotional response and that of the article's protagonist. Additionally, it needs to infer the reader's introspective emotional state - a task that is difficult to achieve with limited contextual information. To illustrate these challenges, the following examples[5] are provided:

1. *Sudden aggression towards brother after medical treatment!*

2. *Climate researchers certain: 1.5 degree limit will be exceeded for decades*

For headline 1, the classifier detects *anger*, failing to distinguish between the protagonist's aggression and the reader's emotional response. For headline 2, the dominant emotion classified is *fear*, which is questionable without additional information about the reader's political orientation or the assessment of the personally perceived threat of climate change. Thus, gaining introspective insights

[5] All example headlines were published in German newspapers and are displayed in their translated version for the purpose of this paper

into the reader's mindset would be crucial for accurately determining the emotion the headline is likely to evoke in the reader. This introspective uncertainty becomes particularly challenging when dealing with polarizing topics such as politics.

Disambiguation. Confusion among the three negatively connoted emotions is notable, which aligns with the quantitative findings reported in Section 4. For illustration, consider the following headlines:

3. *Nurse gave deadly pain patch - suspended sentence*

4. *Smugglers allegedly forced fugitives off speedboat - four dead*

In both cases 3 & 4, the classifier detects *sadness*. However, distinguishing between the negative emotions appears to be challenging, as both *fear* and *anger* could also be relatable emotional responses to these headlines.

Topic matching. The classifier demonstrates a tendency to associate specific topics with particular emotions. For example, political content is predominantly classified as *anger*, reports of personal tragedy and death as *sadness* (see headlines 3 & 4), and report of global catastrophes as *fear*. This suggests the presence of a confounding variable *headline topic*, that likely influences the emotion classification.

6 Conclusion

This paper contributes to research on textual emotion classification, as it (a) provides to the best knowledge of the authors a novel approach for translating and evaluating text corpora by utilizing back-translation and cosine similarity as quality control measure and (b) contributes to the sparse body of research on reader emotion detection in text. (c) By comparing classifier performances on German and English versions of a shared dataset, it provides benchmark results for multilingual emotion classification, demonstrating the superior performance of the English classifier. (d) Using supervised models for emotion classification that combine ADA embeddings with lexical features substantially enhances performance compared to benchmark approaches by approximately 42%. (e) The study empirically supports the significance of consumer emotion in online marketing by highlighting how varying interaction rates for advertisements are influenced by the emotional context of the advertisement.

7 Limitations

Given that emotional responses are latent and highly subjective, attaining unambiguous and objective classification results is hardly achievable. This was illustrated in Section 5.2 and is also supported empirically, as all of the employed datasets report low inter-rater reliability (r = 0.51 over all three datasets). Consequently, the classification accuracy attainable by machine learning algorithms is inherently constrained by the degree of agreement among human annotators.

Furthermore, the imbalance in emotional polarity within the base emotions appeared to affect classification performance at the class level. Thus, these findings could serve as a foundation for future research aimed at developing a more balanced set of base emotions, with a more equitable distribution of emotion polarity.

While this study is confined to using newspaper headlines for predicting user emotion, future research should incorporate additional information. Specifically, future work should (a) analyze the emotional tone of advertisements to explore potential interaction patterns between the emotional appeal of advertisements and that of newspaper headlines, and (b) adopt a multi-modal approach that considers both textual and visual features. Given that images in newspaper articles may convey emotional appeal that either complements or interacts with the emotional message of the text, integrating these visual elements could provide a more comprehensive understanding of emotional impact on the reader.

References

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE.

Nastaran Babanejad, Ameeta Agrawal, Heidar Davoudi, Aijun An, and Manos Papagelis. 2019. Leveraging emotion features in news recommendations. *INRA@ RecSys*, 2554:70–78.

Anil Sriharsha Bandhakavi, Nirmalie Wiratunga, Deepak P, and Stewart Massie. 2016. Lexicon based

feature extraction for emotion text classification. *Pattern Recognition Letters*, 93.

Angelo Basile, Guillermo Pérez-Torró, and Marc Franco-Salvador. 2021. Probabilistic ensembles of zero-and few-shot learning models for emotion classification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 128–137.

Jonah Berger. 2013. *Contagious: Why things catch on*. Simon and schuster.

Jonah Berger and Katherine L Milkman. 2012. What makes online content viral? *Journal of marketing research*, 49(2):192–205.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Laura Bostan, Evgeny Kim, and Roman Klinger. 2019. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. *arXiv preprint arXiv:1912.03184*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, and Amanda Askell. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Lea Canales and Patricio Martínez-Barco. 2014. Emotion detection from text: A survey. In *Proceedings of the workshop on natural language processing in the 5th information systems research working days (JISIC)*, pages 37–43.

Jungsil Choi, Priyamvadha Rangan, and Surendra N Singh. 2016. Do cold images cause coldheartedness? the impact of visual stimuli on the effectiveness of negative emotional charity appeals. *Journal of Advertising*, 45(4):417–426.

Leda Cosmides and John Tooby. 2000. Evolutionary psychology and the emotions. *Handbook of emotions*, 2(2):91–115.

Bao Minh Doan Dang, Laura Oberländer, and Roman Klinger. 2021. Emotion stimulus detection in german news headlines. *arXiv preprint arXiv:2107.12920*.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, et al. 2024. A survey on in-context learning. *Preprint*, arXiv:2301.00234.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & emotion*, 6(3-4):169–200.

Nico H. Frijda. 1986. *The emotions.* Cambridge University Press.

Omar Juarez Gambino and Hiram Calvo. 2019. Predicting emotional reactions to news articles in social networks. *Computer Speech & Language*, 58:280–303.

Raj Kumar Gupta and Yinping Yang. 2018. Crystalfeel at semeval-2018 task 1: Understanding and detecting emotion intensity using affective lexicons. In *Proceedings of the 12th international workshop on semantic evaluation*, pages 256–263.

Jochen Hartmann, Juliana Huppertz, Christina Schamp, and Mark Heitmann. 2019. Comparing automated text classification methods. *International Journal of Research in Marketing*, 36(1):20–38.

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, et al. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Kiana Kheiri and Hamid Karimi. 2023. Sentimentgpt: Exploiting gpt for advanced sentiment analysis and its departure from current machine learning. *arXiv preprint arXiv:2307.10234*.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, et al. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

F. Javier Otamendi and Dolores Lucia Sutil Martín. 2020. The emotional effectiveness of advertisement. *Frontiers in Psychology*, 11.

Robert Plutchik. 1984. Emotions: A general psychoevoiutionary theory. In *Approaches to emotion*, pages 197–219. Psychology Press.

Karolien Poels and Siegfried Dewitte. 2019. The role of emotions in advertising: A call to action. *Journal of Advertising*, 48(1):81–90.

Mehmet Sahin and Nilgun Dungan. 2014. Translation testing and evaluation: A study on methods and needs. *Translation & Interpreting: The International Journal of Translation and Interpreting Research*, 6(2):67–90.

Klaus R Scherer and Harald G Wallbott. 1994. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of personality and social psychology*, 66(2):310.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 70–74.

Judy Strauss, Adel I. Ansary, and Raymond Frost. 2005. *E-marketing*, volume 4. Pearson/Prentice-Hall.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.