Adapting Emotion Detection to Analyze Influence Campaigns on Social Media

Ankita Bhaumik, Andy Bernhardt, Gregorios A Katsios,
Ning Sa, Tomek Strzalkowski
Rensselaer Polytechnic Institute, Troy, New York

Abstract

Social media is an extremely potent tool for influencing public opinion, particularly during important events such as elections, pandemics, and national conflicts. Emotions are a crucial aspect of this influence, but detecting them accurately in the political domain is a significant challenge due to the lack of suitable emotion labels and training datasets. In this paper, we present a generalized approach to emotion detection that can be adapted to the political domain with minimal performance sacrifice. Our approach is designed to be easily integrated into existing models without the need for additional training or fine-tuning. We demonstrate the zero-shot and few-shot performance of our model on the 2017 French presidential elections and propose efficient emotion groupings that would aid in effectively analyzing influence campaigns and agendas on social media.

1 Introduction

Digital environments, such as social media, are powerful launching platforms for wide-reaching influence campaigns surrounding important events such as elections, pandemics, and armed conflicts, as well as commercial interests (Karlsen and Enjolras, 2016; Raudeliūnienė et al., 2018; Badawy et al., 2019). These campaigns aim to manipulate public opinion in a particular way: to favor or oppose a political candidate, to accept or resist vaccination, to justify an aggression, etc. This is achieved by disseminating messages that advance a specific agenda, using language, imagery, and topics that are likely to resonate with the target audience.

Presidential elections offer a substantial context for examining influence campaigns on social media platforms and is the focus of this study. Various indicators, such as agenda, stance, concern, belief, emotion, and imageability, have been identified for measuring the influence of social media messages within this context (Mather et al., 2022).

Emotion is deeply integrated in political discourse and is used as a rhetorical tool in persuading the audience (Cislaru, 2012). Emotionally charged messages can significantly sway public opinion regarding specific agendas or candidates (Weber, 2013; Mohammad et al., 2015) and several studies have documented the effect of emotional language in disseminating polarizing content via social media platforms (Brady et al., 2017).

Existing social media datasets, especially those focused on election-related messages posted on Twitter, are labeled using traditional emotion categories derived from Ekman or Plutchik labels (Ekman, 1999; Plutchik, 1984). These datasets facilitate the development of emotion analysis tools and apply them on diverse applications ranging from healthcare (Tivatansakul et al., 2014) and education (Karan et al., 2022) to stock market (Aslam et al., 2022) and political opinion mining (Cabot et al., 2020). However, each new application domain presents its own set of challenges that existing systems are unable to handle. Therefore, when a new emotion detection problem emerges in a specialized domain, researchers engage in an exhaustive annotation process to build relevant datasets. This highlights the necessity for enhancing the flexibility and robustness of existing models in order to accommodate new scenarios.

Potential solutions involve using semi-supervised, unsupervised, zero-shot, or few-shot techniques (Yin et al., 2019; Chen et al., 2022; Zhang et al., 2019). Nevertheless, solely relying on emotion labels and their definitions from external resources, such as WordNet (Strapparava et al., 2004), are insufficient to capture the intricate concepts and subtleties associated with each emotion label when viewed through the lens of the application domain. Psychological theories suggest that emotion definitions are not universally applicable across domains or individuals, rather, they are profoundly shaped by the socio-cultural

Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 441–451

context and specific events (Averill, 1980; Mohammad and Kiritchenko, 2018), emphasizing the need to incorporate domain-specific knowledge and emotion inter dependencies for effective zero-shot systems.

But which emotions matter in an influence campaign? Do the same emotions arise when discussing a new electronic gadget on the market as when comparing political candidates ahead of an election? In this paper, we present a novel zero-shot approach to detect emotions in text, adaptable to unexplored domains or target label sets. Our method incorporates interpretations of emotion labels and their inter dependencies for improved results in the target domain. We investigate tweets around the 2017 French Presidential Elections part of which is publicly available on Kaggle (Daignan, 2017) and thoroughly evaluate our method to demonstrate that it addresses the shortfalls of existing zero-shot approaches. This is an important step towards providing valuable insights on the emotions of the audience towards political campaigns and agendas.

2 Background

2.1 Emotions in Political Discourse

Extensive research has been conducted on the strategic employment of emotions to sway voting behaviors and public opinion during political campaigns. Campaigns often utilize specific emotional appeals, such as positive emotions (e.g., enthusiasm and pride) to foster support, while leveraging negative emotions (e.g., fear and anger) to incite negative emotions towards the opposition (Ridout and Searles, 2011; Fridkin and Kenney, 2012; Grüning and Schubert, 2022). Some studies contend that only certain emotions, namely anxiety and enthusiasm, are particularly influential in political contexts (Marcus and MacKuen, 1993), with anger and other negative emotions frequently employed by political leaders (Cislaru, 2012).

Prior studies have also reported that negative campaign emotions, such as anger, contempt, disgust, and fear often co-occur and are difficult to distinguish (Fridkin and Kenney, 2012; Mohammad and Kiritchenko, 2018). Consequently, the selection of emotion labels is heavily reliant on the specific influence patterns under examination, which presents the challenge of developing a versatile emotion model capable of adapting to various emotion label sets.

2.2 Emotion Detection Models

Emotion detection in text is a long-standing research challenge due to the ever-changing nature of textual content across applications and platforms. Large pretrained language models, such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019), have emerged as powerful tools for this task (Cai and Hao, 2018; Huang et al., 2019; Polignano et al., 2019; Ma et al., 2019; Chiorrini et al., 2021). Our approach employs popular Twitter-specific language models, which provide a robust baseline for core NLP tasks in social media analysis (Barbieri et al., 2020).

Zero-shot learning techniques are frequently employed for emotion detection when training data is unavailable in the target domain. Recent studies in zero-shot emotion detection use text entailment approaches, wherein target labels generate hypotheses for the model (Yin et al., 2019; Basile et al., 2021). Prompt engineering techniques also facilitate emotion label inference from pretrained NLI models (Plaza-del Arco et al., 2022). Additionally, some zero-shot methods leverage sentence embeddings for unsupervised or semi-supervised predictions on unlabeled datasets (Chen et al., 2022; Zhang et al., 2019; Olah et al., 2021). The drawback of these techniques stem from their generalized design, enabling them to function across multiple domains, while only excel when target emotion labels align with standard definitions. They lack integration of domain knowledge or comprehension of emotion concepts that may arise in specialized domains.

3 Methodology

Upon completing a preliminary investigation of the 2017 French election dataset, our annotation team observed that assigning a distinct emotion label to each tweet is a challenging and a complex task. Following a more practical approach, we label tweets using groups of emotions that frequently co-occur or overlap (Mohammad and Kiritchenko, 2018; Cislaru, 2012). These groups of emotions are combinations of the traditional emotion labels and are difficult to isolate from short informal tweets. In instances where a message cannot be classified into any of the emotion groups but still conveys a strong positive or negative sentiment, it is assigned a "Positive-other" or "Negative-other" label. The following is the final set E of grouped emotion labels:

1. Anger, hate, contempt, disgust
2. Embarrassment, guilt, shame, sadness
3. Admiration, love
4. Optimism, hope
5. Joy, happiness
6. Pride, including national pride
7. Fear, pessimism
8. Sarcasm, amusement
9. Positive-other
10. Negative-other

3.1 Problem Statement

The goal is to automatically tag a text message x with scores between 0 and 1 for each emotion label in E. The score for each label $e \in E$ should reflect the confidence that the emotion e is expressed by the author of x.

3.2 Approach

Our approach combines traditional sentiment analysis and emotion detection results, obtained by hierarchical grouping of standard emotions. The grouped emotion labels form the basis for our ensemble model, which can be readily adapted to the political domain without the need for additional training. The performance of this ensemble can be further optimized with the availability of some in-domain annotated data. We illustrate our emotion model ensemble in Fig.1.

Given the text of a tweet as input, our model produces scores over three sentiment categories, six Ekman emotions, and their fine-grained subcategories defined in (Demszky et al., 2020). To obtain confidence scores over label set E, we design a many-to-one mapping based on the emotion groups and their corresponding definitions in the political domain.

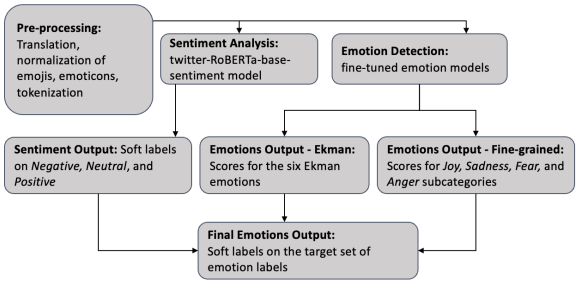


Figure 1: Ensemble Emotion Detection Architecture

3.3 Datasets & Preprocessing

We have identified two social media datasets that can be utilized to train the emotion models in our ensemble, providing us with the broadest possible coverage of all standard and fine-grained emotion labels:

Cleaned Balanced Emotional Tweets (CBET) (Shahraki and Zaiane, 2017) is a collection of 81k English tweets that have been collected using a set of hashtags corresponding to the nine emotion labels (*anger, fear, joy, love, sadness, surprise, thankfulness, disgust*, and *guilt*). We use this dataset to train a model to predict scores over the six Ekman emotions, removing the annotations for *thankfulness, love* and *guilt*. The 56,281 remaining tweets that have at least one emotion label are split randomly into training (81%), validation (9%), and testing (10%) sets.

GoEmotions (Demszky et al., 2020) is a corpus of 58k English Reddit comments manually annotated with 27 emotion labels or *neutral*. The large number of fine-grained emotion labels in this dataset makes it an ideal choice for creating a base emotion model suitable specialized emotion tasks. We use GoEmotions to train a model to predict scores over the six Ekman emotions, and for the emotions of *joy, sadness, fear* and *anger*, we identify their lower level emotions in the hierarchy of the dataset to produce the training, validation, and testing sets (Table 1) to train specialized emotion models.

Model	Training	Validation	Test
joy	17,410	2,219	2,104
sadness	3,263	390	379
fear	726	105	98
anger	5,579	717	726

Table 1: Distribution of training, validation, and test sets for emotion subcategory models derived from GoEmotions

Given an input tweet, our system first translates it to English [1] and applies basic text preprocessing techniques (Tiedemann and Thottingal, 2020). The preprocessing pipeline is used as a social tokenizer (Baziotis et al., 2017) to remove any usernames, tweet IDs, hyperlinks, emails, phone numbers, times, dates, and percentages, normalize money

[1] https://huggingface.co/Helsinki-NLP/opus-mt-fr-en

values and numbers, annotate any censored or elongated words, and convert emoticons to plain text.

3.4 Training and Fine-tuning

For the task of sentiment analysis, we use the twitter-XLM-RoBERTa-base-sentiment [2] model to produce normalized values on the three sentiment categories *negative, neutral*, and *positive* (Barbieri et al., 2020).

For emotion detection, we further fine-tune six models as components of the hierarchical mapping system. Each model in the ensemble is built using the twitter-RoBERTa-base-emotion [3] (Barbieri et al., 2020) checkpoint, but we append a new linear layer on top of the last hidden state of RoBERTa's [CLS] token. The purpose of the linear layer is to convert the final hidden state vector into a vector related to the distinct emotion labels in the corresponding dataset. Subsequently, this vector can be converted into probabilities via the Softmax function. The labels of each model are listed in Table 2.

In the first step, two models are fine-tuned to output normalized scores on the six Ekman emotions using the CBET Twitter data and GoEmotions Reddit data. We choose to train separate models on Twitter and Reddit data to be able to weigh them in the next step based on the target domain. The remaining four models are then fine-tuned to output scores on the subcategories of *joy, sadness, fear, and anger*. The fine-tuning setup and metrics for each model are described in Appendix A.

3.5 Hierarchical Label Transfer

The fine-grained emotion scores are used downstream to adapt the model to a new domain. We map the scores from the model outputs to scores over a desired label set E using a weighted linear combination derived by considering the relatedness of emotions in the Plutchik's wheel of emotions (Plutchik, 1984) and the co-occurrence of emotions in the target domain. A general set of rules to determine the mapping from the hierarchical emotion model outputs to other emotions $e \in E$ is outlined below:

1. Determine which sentiment category (positive/negative) $S \subseteq Sent$ corresponds to emotion e. (e.g. *e: optimisms* $\rightarrow$ *s: positive*). For ambiguous emotions, we choose the sentiment category with a higher score.

2. To remove any bias caused by a specific dataset, calculate one output score EK for each Ekman label using a linear combination of the scores from the CBET-Ekman and GE-Ekman models.

3. For each sentiment $s \in S$, determine which high-level Ekman emotions corresponding to s, $EK_s \subseteq EK$ have subcategories relevant to emotion e. As mentioned above, the sentiment of *e: optimism* is *positive*, and *joy* is the EK_s which corresponds to *s: positive* and has subcategories relevant to *e: optimism*.

4. For each high-level Ekman emotion $ek \in EK_s$, if ek has subcategories, determine which subcategories $sub_{ek} \subseteq Sub_{ek}$ are relevant to emotion e. Continuing with the example of *optimism*, out of all the *Joy* subcategories, the only relevant subcategory is *optimism*.

5. Then, the score of e is

$$\sum_{s \in S} \sum_{ek \in EK_s} \sum_{sub_{ek} \in Sub_{ek}} \alpha \left(Sent[s] * EK[ek] * Sub_{ek}[sub_{ek}] \right)$$

where α is a weight that can be set to 1, or fine-tuned to maximize a performance metric on a target-domain validation set (if one exists). In other words, the final score for e is a weighted sum of terms, where each term

Model	Output Labels
Sentiment(Sent)	positive, neutral, negative
CBET-Ekman	joy, sadness, fear, anger, disgust, surprise
GE-Ekman	joy, sadness, fear, anger, disgust, surprise
Joy(J)	joy, amusement, approval, excitement, gratitude, love, optimism, relief, pride, admiration, desire, caring
Sadness(S)	sadness, disappointment, embarrassment, grief, remorse
Fear(F)	fear, nervousness
Anger(A)	anger, annoyance, disapproval

Table 2: Set of output labels for each component model.

[2] https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment

[3] https://huggingface.co/cardiffnlp/twitter-roberta-base-emotion

Mapping	Output Label
((EK[anger] * A[anger]) + EK[disgust]) * Sent[negative]	anger, contempt, disgust
(EK[sadness] * (S[sadness] + S[embarrassment] + Sent[grief])) * Sent[negative]	embarrassment, guilt
(EK[joy] * (J[admiration] + J[love])) * Sent[positive]	admiration, love
(EK[joy] * (J[optimism])) * Sent[positive]	optimism, hope
(EK[joy] * (J[joy])) * Sent[positive]	joy, happiness
(EK[joy] * (J[pride])) * Sent[positive]	pride
(EK[fear] * (F[fear])) * Sent[negative]	fear, pessimism
(EK[joy] * (J[amusement])) * Sent[positive]	amusement, sarcasm
(EK[joy] * (J[approval] + J[excitement] + J[gratitude] + J[relief] + J[desire] + J[caring])) * Sent[positive]	positive-other
((EK[sadness] * (S[disappointment] + S[remorse])) + (EK[fear] * (F[nervousness])) + (EK[anger] * (A[annoyance] + A[disapproval]))) * Sent[negative]	negative-other

Table 3: Mapping of model outputs to French election labels

is the product of scores for a sentiment, Ekman emotion, and low-level emotion subcategory triple that is relevant to e. For example, for the output emotion *optimism*, we may have the term $(Sent[positive] * EK[joy] * Joy[optimism])$.

3.6 In-domain Optimization

The availability of any in-domain data can be used as a validation set to boost the model performance in two ways: 1) finding a set of optimal classification thresholds for each emotion label, and 2) fine-tuning the weights of the linear mapping of the emotion scores for a target metric. The classification thresholds are fine-tuned by choosing a threshold for each target class to maximize the F1 score on that class over the validation dataset.

The mapping weights are optimized by successively applying differential evolution to each individual target label mapping to maximize the F1 score on that label over the validation dataset (Storn and Price, 1997). We use a subset of the manually annotated French election dataset to fine-tune both the mapping weights and the classification thresholds by first optimizing the weights, and subsequently choosing the thresholds for each label. More details on the label-wise classification thresholds and mapping weights parameters have been listed in Appendix B.

3.7 Data

Our annotation team utilized the emotion label set E, as detailed in Section 3, to annotate a subset of the 2017 French Presidential Election dataset. Three raters independently assigned one or more emotions to each tweet, with a label considered ground truth if confirmed by at least two annotators. The inter-rater reliability (IRR) across all emotion labels for the three raters was determined to be 17%, calculated by macro-averaging kappa scores (Carletta, 1996) between each rater pair. This low IRR highlights the task's complexity and the challenge of obtaining consistent emotion labels in this domain. Factors such as political background familiarity, cross-cultural dynamics, and multilingualism contribute to this complexity (Shaikh et al., 2015).

In addition, the annotators assigned agenda labels as a second influence indicator to the dataset. An agenda can be defined as the indicator that influences the target audience to believe in something or to perform a certain task (e.g., vote for a candidate, engage in a demonstration). We perform a set of experiments that utilize these agenda labels to examine the emotional patterns in relation to different agendas in a campaign. We show that the use of emotional language tends to align strongly with the type of agenda being promoted.

4 Experiments

In this section, we compare our approach against popular semi-supervised and zero-shot techniques. All experiments have been carried out on the French election dataset in the below configurations:

- *Zero-shot mode:* Emotion classification on the test set by adapting the model ensemble to the

target domain without any fine-tuning. We also repeat this experiment without the sentiment component in the ensemble to demonstrate its contribution.

- *In-domain optimization mode:* Use a small subset of available in-domain data to optimize the classification thresholds and mapping weights.

4.1 Baselines

We evaluate our approach against the following baselines:

- *Zero-shot textual entailment (TE):* Following the work of Yin et al., 2019, we convert each emotion label into the hypothesis: "This text expresses <label>." We use the BART MNLI [4] model to generate entailment and contradiction scores and threshold them to produce binary outputs for each label.
- *Zero-shot sentence embeddings (SB):* We use SBERT (Reimers and Gurevych, 2019) to obtain the embeddings for the input texts and class labels [5]. The texts are then labeled based on their closeness to the labels in the embedding space using cosine similarity.
- *Semi-supervised models (EK):* We use existing emotion datasets (CBET and GoEmotions) to fine-tune twitter-RoBERTa-base-emotion pretrained models (Barbieri et al., 2020) on the six Ekman labels, and test these models over the label set in the target domain. Many of the target labels are absent in these Ekman datasets and thus their outputs are set to 0.

4.2 Results

The mapping of the model ensemble outputs to the French election emotion groupings (Table 3) follows the general rules outlined in Section 3.4. For example, each label in *anger, hate, contempt, disgust* is associated with a *negative* sentiment. Further, for the Ekman emotions *anger* and *disgust*, the only relevant subcategory is *anger*, which results in the final mapping *((EK[anger] * Anger[anger]) + EK[disgust]) * Sentiment[negative]*. The label *positive-other* is associated with a *positive* sentiment and the only positive Ekman emotion, *joy*. Additionally, from the label definition, it accumulates scores of all the positive fine-grained emotions that have not been recorded by any other label. Figure 2 shows an example tweet from the test dataset with its output emotion scores.

The evaluation metrics in Table 4 highlight the poor performance of existing zero-shot methods on the French Election dataset. This is because these models do not incorporate any domain knowledge and rely solely on the standard emotion definitions to classify text. The specialized label groups in the French election labels are tailored to the influence detection task, which makes them difficult for traditional emotion detection models to handle. For example, the labels *love* and *admiration* can be synonymous in a political influence campaign but not in a general emotion taxonomy. This further emphasizes the need for domain-specific knowledge in emotion detection models which is incorporated by our label transfer step.

	EK	TE	SB	Ours
anger/cont/disgust	0.17	0.13	0.13	0.23
embarrass/guilt	0.05	0.03	0.04	0.19
admiration/love	0	0.04	0.04	0.15
optimism/hope	0	0.22	0.16	0.30
joy/happiness	0.04	0.04	0.03	0.16
pride	0	0.07	0.07	0.17
fear/pessimism	0.10	0.07	0.06	0.18
amusement	0	0.14	0.14	0.14
positive-other	0	0.56	0.43	0.50
negative-other	0	0.53	0.41	0.50

Table 4: F1 scores across all emotion labels in the French Election dataset; (Ours: zero-shot performance of deploying our emotion model ensemble over this dataset)

For the few-shot mode (Table 5), the optimization of the classification thresholds and label mapping weights do not aid in improving the performance of the model due to inconsistencies in annotation between the validation and test datasets. We believe that more consistent annotations or sampling fine-tuning data from the same dataset would result in a performance boost as observed in other datasets performing the same task.

We also perform an ablation study to understand the effect of adding a sentiment component to the ensemble (Table 5). The improvement of scores across all experiments ascertain that the influence of sentiment is crucial for this emotion detection task.

[4] https://huggingface.co/facebook/bart-large-mnli

[5] We use the SBERT all-MiniLM-L6-v2 pretrained model to obtain the embeddings.

	P	R	F1
Semi-supervised			
CBET	0.05	0.07	0.06
GoEmotions	0.05	0.08	0.06
CBET + GoEmotions	0.06	0.09	0.07
Zero-shot			
BART MNLI (TE)	0.13	0.86	0.23
SBERT (SB)	0.10	0.65	0.17
Ours	0.32	0.44	0.37
Ours + Sentiment	0.34	0.48	**0.40**
Few-shot			
Optimize mapping	0.34	0.48	**0.39**
Optimize threshold	0.29	0.29	0.29

Table 5: Evaluation results against baselines. Ours: Our emotion model ensemble without the sentiment module; Ours + Sentiment: Our emotion model ensemble with sentiment module. *Few-shot* section lists results of optimizing our label transfer step with the availability of some in-domain data.

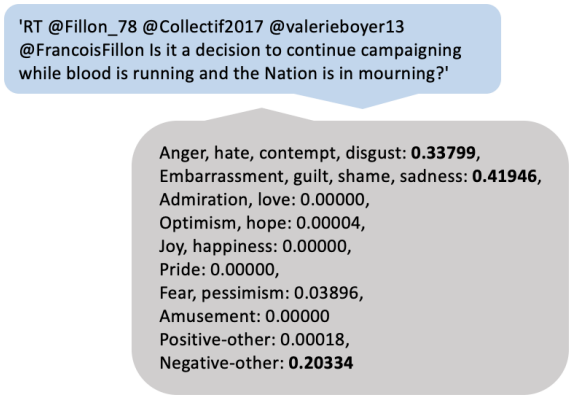


Figure 2: Example tweet from the French election dataset.

5 Emotions as an Influence Indicator

We use our emotion detection approach to understand how emotions correlate with other influence indicators during political campaigns. We select a subset of the tweets that are associated with specific agendas in the election. Figure 5a shows the emotion distribution across tweets mentioning popular candidates. As expected, the predominant emotions are *anger, contempt, disgust* and *optimism, hope* signifying that political campaigns either influence the audience by expressing hope/optimism for a brighter future or by expressing hatred towards the opposing candidate or political party. Interestingly, although a large portion of the tweets express some strong positive/negative emotion, they cannot be accurately tagged with a specific emotion label (Fig 3). This leads us to hypothesize that a large number of emotion labels may not be required to effectively analyze the emotional influence of political campaigns.

Figures 4 and 5b illustrates the results of emotion detection on the agenda annotated tweets. In this paper, we focus on the following agendas: 1) believe that an entity (E) or group (G) is immoral or harmful; 2) believe that E/G is moral or beneficial; 3) believe that your group are at risk; 4) believe that your actions can lead to a good outcome or hope; 5) call to share information; 6) call to vote for a E/G; 7) call to vote against a E/G; and 8) call to participate in demonstration/protest or attend a rally/campaign.

In Figure 4, higher anger and negative-other scores are observed in the agenda of 'Entity is immoral'. In contrast, the 'Entity is moral' agenda shows higher *admiration* and *positive-other* scores. By comparing the emotions of 'group at risk' and 'belief for good outcome', we find higher *anger, fear*, and *negative-other* scores in the former, and higher *optimism* and *positive-other* scores in the latter. Similarly, the agendas 'vote for entity' and 'vote against entity' have higher positive and negative emotions, respectively. These differences can be clearly seen in Fig. 5b that shows the proportion of each emotion in an agenda. We can conclude that emotions play an important role in understanding patterns in a campaign and the impact of political agendas on the audience.

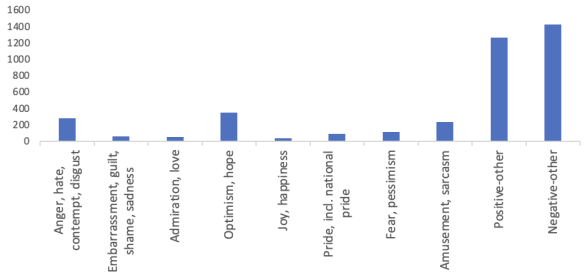


Figure 3: Distribution of emotion labels in the test dataset

6 Conclusion

Our paper presents a generalized approach to emotion detection wherein existing emotion detection datasets and models can be quickly adapted to specialized emotion labels to effectively analyze influence campaigns in the political domain. Our experiments demonstrate the efficacy of this zero-shot approach on tweets from the 2017 French presidential election.

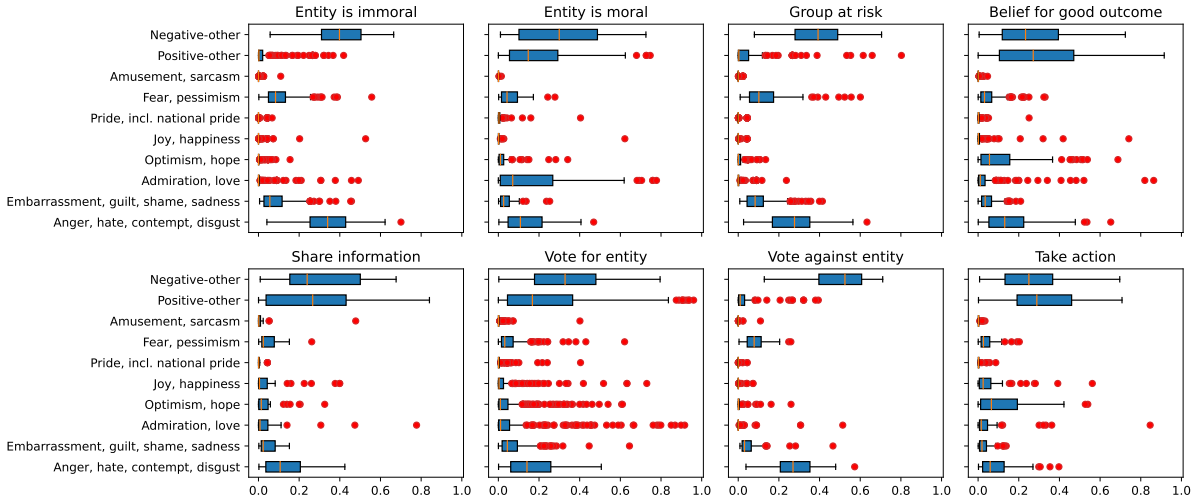
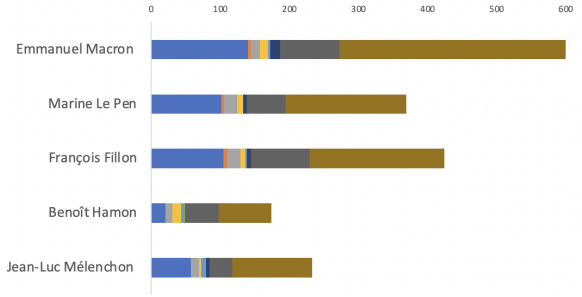
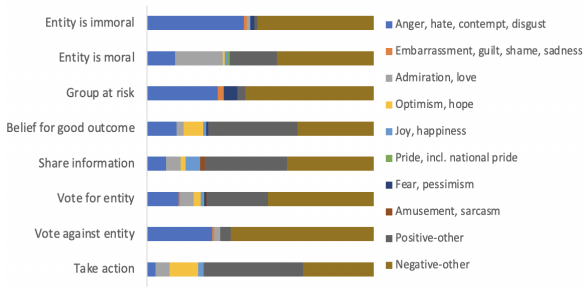


Figure 4: Boxplots showing summary of emotion scores across different agendas in the campaign. The box from the first quartile to the third quartile, the line depicting the median score for that label. The whiskers are shown from the box by 1.5x the inter-quartile range. Anything past the whiskers are shown as outliers in red.



(a) Distribution of emotion labels across tweets mentioning popular candidates



(b) Distribution of emotion labels across different agendas in the campaign

Figure 5: Distribution of emotions during campaigns for political figures or agendas

We further utilize our inference results to get insights on the use of emotional language along with other influence indicators like agenda. This work could be utilized in multiple downstream applications to forecast election outcomes or understand public opinions on specific agendas or issues. Our results signify the importance of certain emotion groups in political campaigns and provides a path for future work integrating multiple influence indicators in social media and understanding interdependencies between different emotions.

7 Limitations

Currently our approach relies on translation to analyze multilingual tweets. Future work would include using multilingual pre-trained models like XLM-RoBERTa and the use of non-English training data to build a language agnostic emotion model ensemble.

We carry out our in-domain optimization on a small validation dataset that was annotated by a different set of raters than the one used for the test dataset, which results in a performance drop in the few-shot mode. Ideally, the availability of a high quality validation dataset would boost the zero-shot performance and further adapt the label mappings to the target domain. We also aim to carry out in house annotations by experts to release a publicly available dataset annotated with emotions in the political domain which would pave the way for further analysis in this domain.

Acknowledgements

This paper is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. HR001121C0186. Any opinions, findings and conclusions or recommendations expressed in this material are those of

the authors and do not necessarily reflect the views of DARPA or the U.S. Government.

Ethics Statement

We use multiple Twitter and Reddit datasets to fine-tune our emotion model ensemble. Both these datasets have been cleaned to remove any toxicity, biases and offensive language. The annotated French election dataset cannot be publicly released following the terms and conditions of the project. The data available to us for fine-tuning and evaluation does not contain any personally identifiable data and we do not have any knowledge of the annotators behind creating this dataset. We also utilize multiple pre-trained models which reduces the carbon footprint of training models from scratch. Further, utilization of this transfer learning method for any new domain would not incur any training costs as minimal fine-tuning may be required. However, the results obtained in an unknown domain should be human evaluated before using it for any downstream analytics task.

References

Naila Aslam, Furqan Rustam, Ernesto Lee, Patrick Bernard Washington, and Imran Ashraf. 2022. Sentiment analysis and emotion detection on cryptocurrency related Tweets using ensemble LSTM-GRU Model. *IEEE Access*, 10:39313–39324.

James R Averill. 1980. A constructivist view of emotion. In *Theories of emotion*, pages 305–339. Elsevier.

Adam Badawy, Aseel Addawood, Kristina Lerman, and Emilio Ferrara. 2019. Characterizing the 2016 russian ira influence campaign. *Social Network Analysis and Mining*, 9:1–11.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650.

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*, pages 128–137.

Christos Baziotis, Nikos Pelekis, and Christos Doulkeridis. 2017. DataStories at SemEval-2017 task 4: Deep LSTM with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

Pere-Lluís Huguet Cabot, Verna Dankers, David Abadi, Agneta Fischer, and Ekaterina Shutova. 2020. The pragmatics behind politics: Modelling metaphor, framing and emotion in political discourse. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4479–4488.

Xiaofeng Cai and Zhifeng Hao. 2018. Multi-view and attention-based bi-LSTM for Weibo emotion recognition. In *2018 International Conference on Network, Communication, Computer Engineering*, pages 772–779.

Jean Carletta. 1996. Assessing agreement on classification tasks: The kappa statistic. *Computational Linguistics*, 22(2):249–254.

Qi Chen, Wei Wang, Kaizhu Huang, and Frans Coenen. 2022. Zero-shot text classification via knowledge graph embedding for social media data. *IEEE Internet of Things Journal*, 9(12):9205–9213.

Andrea Chiorrini, Claudia Diamantini, Alex Mircoli, and Domenico Potena. 2021. Emotion and sentiment analysis of tweets using BERT. In *EDBT/ICDT Workshops*.

Georgeta Cislaru. 2012. Emotions as a rhetorical tool in political discourse. In Maria Zaleska, editor, *Rhetoric and Politics*, pages 107–126. Cambrige Scholar Press.

Jean-Michel Daignan. 2017. French presidential election: Extract from twitter about the french election.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Paul Ekman. 1999. Basic emotions. *Handbook of cognition and emotion*, 98(45-60):16.

Kim L Fridkin and Patrick J Kenney. 2012. The impact of negative campaigning on citizens' actions and attitudes. *The SAGE handbook of political communication*, pages 173–185.

David J Grüning and Thomas W Schubert. 2022. Emotional campaigning in politics: Being moved and anger in political ads motivate to support candidate and party. *Frontiers in Psychology*, 12:6337.

Yen-Hao Huang, Ssu-Rui Lee, Mau-Yun Ma, Yi-Hsin Chen, Ya-Wen Yu, and Yi-Shin Chen. 2019. EmotionX-IDEA: Emotion BERT- an Affectional Model for Conversation. *arXiv preprint arXiv:1908.06264*.

KV Karan, Vedant Bahel, R Ranjana, and T Subha. 2022. Transfer learning approach for analyzing attentiveness of students in an online classroom environment with emotion detection. In *Innovations in Computational Intelligence and Computer Vision: Proceedings of ICICV 2021*, pages 253–261. Springer.

Rune Karlsen and Bernard Enjolras. 2016. Styles of social media campaigning and influence in a hybrid political communication system: Linking candidate survey data with twitter data. *The International Journal of Press/Politics*, 21(3):338–357.

Luyao Ma, Long Zhang, Wei Ye, and Wenhui Hu. 2019. PKUSE at SemEval-2019 task 3: emotion detection with emotion-oriented neural attention network. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 287–291.

George E Marcus and Michael B MacKuen. 1993. Anxiety, enthusiasm, and the vote: The emotional underpinnings of learning and involvement during presidential campaigns. *American Political Science Review*, 87(3):672–685.

Brodie Mather, Bonnie Dorr, Adam Dalton, William de Beaumont, Owen Rambow, and Sonja Schmer-Galunder. 2022. From stance to concern: Adaptation of propositional analysis to new tasks and domains. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3354–3367.

Saif Mohammad and Svetlana Kiritchenko. 2018. Understanding emotions: A dataset of tweets to study interactions between affect categories. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. 2015. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499.

Justin Olah, Sabyasachee Baruah, Digbalay Bose, and Shrikanth Narayanan. 2021. Cross domain emotion recognition using few shot knowledge transfer. *arXiv preprint arXiv:2110.05021*.

Flor Miriam Plaza-del Arco, María-Teresa Martín-Valdivia, and Roman Klinger. 2022. Natural language inference prompts for zero-shot emotion classification in text across corpora. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6805–6817.

Robert Plutchik. 1984. Emotions: A general psychoevolutionary theory. *Approaches to emotion*, 1984(197-219):2–4.

Marco Polignano, Pierpaolo Basile, Marco de Gemmis, and Giovanni Semeraro. 2019. A comparison of word-embeddings in emotion detection from text using biLSTM, CNN and Self-attention. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 63–68.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding with unsupervised learning.

Jurgita Raudeliūnienė, Vida Davidavičienė, Manuela Tvaronavičienė, and Laimonas Jonuška. 2018. Evaluation of advertising campaigns on social media networks. *Sustainability*, 10(4):973.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Travis N Ridout and Kathleen Searles. 2011. It's my campaign i'll cry if i want to: How and when campaigns use emotional appeals. *Political Psychology*, 32(3):439–458.

Ameneh Gholipour Shahraki and Osmar R Zaiane. 2017. Lexical and learning-based emotion mining from text. In *Proceedings of the international conference on computational linguistics and intelligent text processing*, volume 9, pages 24–55.

Samira Shaikh, Tomek Strzalkowski, Sarah Taylor, John Lien, Ting Liu, George Aaron Broadwell, Laurie Feldman, Boris Yamrom, Kit Cho, and Yuliya Peshkova. 2015. Understanding cultural conflicts using metaphors and sociolinguistic measures of influence. In *Proceedings of the Third Workshop on Metaphor in NLP*, pages 67–76.

Rainer Storn and Kenneth Price. 1997. Differential evolution - a simple and efficient heuristic for global optimization over continuous spaces. *Journal of Global Optimization*, 11:341–359.

Carlo Strapparava, Alessandro Valitutti, et al. 2004. WordNet Affect: An affective extension of wordnet. In *Lrec*, volume 4, page 40. Lisbon, Portugal.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT — Building open translation services for the World. In *Proceedings of the 22nd Annual Conferenec of the European Association for Machine Translation (EAMT)*, Lisbon, Portugal.

Somchanok Tivatansakul, Michiko Ohkura, Supadchaya Puangpontip, and Tiranee Achalakul. 2014. Emotional healthcare system: Emotion detection by facial expressions using japanese database. In *2014 6th computer science and electronic engineering conference (CEEC)*, pages 41–46. IEEE.

Christopher Weber. 2013. Emotions, campaigns, and political participation. *Political Research Quarterly*, 66(2):414–428.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 3914–3923.

Jingqing Zhang, Piyawat Lertvittayakumjorn, and Yike Guo. 2019. Integrating semantic knowledge to tackle zero-shot text classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1031–1040.

A Hyperparameters

To fine-tune the pretrained twitter-RoBERTa-base-emotion models on each of the six training and validation datasets, we use the following settings, chosen in order to stay close to the pretrained weights and also alleviate overfitting to the target domains. We use a binary cross-entropy loss for the task of multi-label classification, an Adam optimizer, an initial learning rate of 1e-6, and a batch size of 16. During each training procedure, we apply early stopping on the validation loss with a patience of 10 epochs to alleviate overfitting by stopping fine-tuning when the validation performance no longer improves. In each case, we choose the model that achieves the lowest validation loss as our final model. We train for 72 epochs on the CBET dataset over the six Ekman emotions, 90 epochs on the GoEmotions dataset over the six Ekman emotions, 66 epochs on the GoEmotions *joy* subcategory dataset, 13 epochs on the GoEmotions *sadness* subcategory dataset, 18 epochs on the GoEmotions *fear* subcategory dataset, and 8 epochs on the GoEmotions *anger* subcategory dataset, in order to achieve these best results in Table 6. Across the six models, the total training procedure converged after approximately 5.5 hours on a single GPU.

B Fine-Tuning Thresholds and Weights

In the hierarchical label mappings in Tables 3, the weights for each term in the linear combinations for each target emotion are by default set to 1. Without any fine-tuning data in the target domain, we let each emotion subcategory have equal weight in determining the value of the target emotion. Additionally, in the evaluation, we let the thresholds for classification of each emotion all be equal to 0.3. However, with the availability of a small in-domain validation dataset, we can improve the classification thresholds as well as the mapping weights. We fine-tune the classification thresholds by choosing a threshold for each target class to maximize the F1 score on that class over the validation dataset.

Model	**Validation Accuracy**	**Test Accuracy**
CBET-Ekman	0.6558	0.6483
GoEmo-Ekman	0.6966	0.6914
Joy	0.7386	0.7519
Sadness	0.7205	0.7625
Fear	0.9048	0.8878
Anger	0.6541	0.6501

Table 6: Final validation accuracy and final testing accuracy for each of the six fine-tuned twitter-RoBERTa-base-emotion models in our model ensemble

We fine-tune the mapping weights by successively applying differential evolution to each individual target label mapping to maximize the F1 score on that label over the validation dataset (Storn and Price, 1997). The implementation of the differential evolution algorithm for fine-tuning the mapping weights is provided by Scipy[6]. For each target label mapping, we constrain each weight in $[0, 2]$ in the optimization process, and continue iteratively until the improvements in the label-wise F1 scores are sufficiently small.

[6] https://docs.scipy.org/doc/scipy/reference/generated/scipy.optimize.differential_evolution.html