# Introduction:

Content analysis has always been one of the key methods in communication research. However, with the advent of web 3.0, the data sources for communication and political content analysis has exponentially grown in volume and complexity. The advances in computational methods, such as supervised machine learning, allow researcher to process vast quantities of text.

Yet, communication rarely happens via a single modality in legacy and contemporary communication channels. For example, one of the key political actors in Europe, the European Union has increased its Twitter communication almost 4-folds in the last decade where more than 40% of the messages contain at lease one embedded image (Anonymized 2022). Dictionary based and shallow learning (SL) methods, unfortunately, have a hard time incorporating multi-modality into the analysis.

Deep learning (DL) brings the possiblity to extend content analysis to multi-modal materials. Previous studies have demonstrated the flexibility of embeddings to analyze multimodel data (K. Li et al., 2022; Niu et al., 2019; Tseng et al., 2021).

In this paper, we evaluate the feasibility of using multi-modal DL embeddings to classify political messages where the message is delivered with a combination of visual and textual modalities in a computational experiment. We build a series of unimodal SL models and multi-modal DL embedding-based models to classify manually annotated tweets from EU executives. We then compare the classification performance of these models. Our results indicate that multimodal signals are tricky to catch in a way that is meaningful to a classifier. Finally, we conclude with some recommendations for researchers who would like to use multi-modal data in automated content analysis.

# Research design and data:

We use manually annotated 898 tweets from the EU executive accounts sent out between December 1st 2019 and July 31st 2020. We focus on these tweets in our experiment specifically because they deliver their message by combining text and images or supplemented the textual message with imagery. Examples of them is presented in Figure 1. Tweets were manually coded as a whole based on whether it provides an update and information on political operations, policies, programs, and reports published by the EU executive institutions and bureaucrats. Overall there are 479 tweets that contain such message (code:1) and 419 tweets that do not (code:0). To ensure the data quality, three rounds of intercoder reliability tests were conducted between coders before coding the full sample(Krippendorf’s α >.8).

Graphical user interface, text

Description automatically generated

Figure 1: Example tweets

In our experiments, we take the binary indicator as the outcome and learn individual SL and DL predictive models. From the SL classification family, we use Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), logistic regression, and XGBoost (XGB) while we use a Multi-Layer Perceptron (MLP) as a DL classifier. We extensively use Quanteda (Benoit et al., 2018) for text preprocessing where we removed links and punctuation. For learning SL models we used caRet (Kuhn, 2008) package in R and SciKitLearn(Pedregosa et al., 2011) in python. For DL models, we utilize the transformers (Wolf et al., 2020) and Keras (Chollet, 2015) libraries in Python.

In our experimental setting, we test the predictive capacity of these models with three different ways of featurization. The first mode of featurization is a document feature matrix of the tweet texts with no weighting. The second is a term-frequency inverse-document-frequency (tf-idf) representation of tweet texts. Thirdly we combine textual and visual elements in a [898, 612, 768] embedding matrix as input data. This means, that for multimodal representations, there is a 612\*768 feature matrix to represent every tweet and its image. of e

This feature matrix is created using VisualBERT (L. H. Li et al., 2019). Figure 2 shows how VisualBERT combines textual and visual input into a combined feature matrix. The textual part is a BERT model, where word embeddings are generated from tokens. The visual embeddings are generated using a pre-trained image model, that detects objects in images. In our case, we use Facebook’s detectron2 library. (Wu et al., 2019) As seen in Figure 2 the image model produces region proposals and transforms these into embedding representations of those regions. Textual and visual embeddings are then passed into a transformer which is again of the same architecture as the BERT base version as proposed by Devlin et al. (2019).

Training of VisualBERT happens by two tasks also closely related to the BERT training procedure. The authors call them (1) Masked language modeling with the image and (2) Sentence-image prediction. In (1) some textual elements from the text input are masked and the model must predict what that text should be. Image input is never masked. In (2) the model must decide which of the two given captions belong to an image. This procedure is “allowing the model to implicitly discover useful alignments between both sets of inputs, and build up a new joint representation.” (L. H. Li et al., 2019, p. 4)

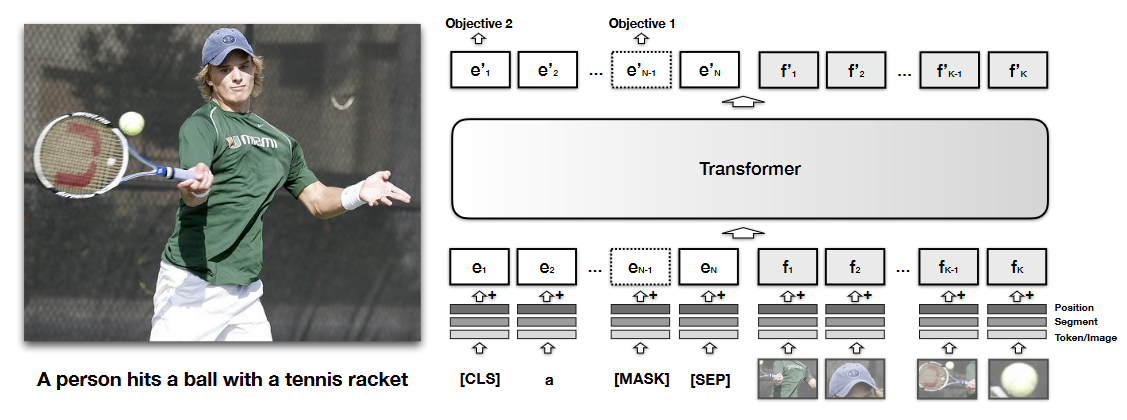


Figure 2: VisualBERT architecture, from L. H. Li et al. (2019)

# Results

We report our experimental results on best performing models with different featurization of text and image data in Table 1. Our experiments show three key results. First and foremost, text only representations with SL models seem to outperform DL models with multi-modal embeddings by a large margin. Our best performing model is a random forest using tf-idf representation of text only data. This is followed by random forest using document-feature matrix. Our second key result is that SL models tend to perform better in terms of precision, but not in recall, when the dataset is text only. Only exception to this is our SVM with tf-idf input where recall outweighs the precision by .10. We do not observe this pattern when the input dataset is multi-modal embedding. Our last key result is multi-layer perceptron tends to outperform other SL learners when data is represented by multi-modal embeddings by about a margin of .10 in F1 score.

Table 1: Performance scores of predictive models

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | DFM | | | TFIDF | | | Multi-modal embeddings | | |
|  | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall |
| Naïve Bayes | 0.57 | 0.48 | 0.70 | 0.37 | 0.51 | 0.29 | - | - | - |
| Logistic regression | 0.66 | 0.68 | 0.64 | 0.70 | 0.70 | 0.71 | - | - | - |
| Support Vector machine | 0.70 | 0.71 | 0.69 | 0.77 | 0.73 | 0.82 | 0.53 | 0.53 | 0.53 |
| Random forest | **0.72** | 0.75 | 0.69 | **0.78** | 0.78 | 0.77 | 0.52 | 0.52 | 0.52 |
| XGBoost | 0.67 | 0.64 | 0.70 | 0.71 | 0.72 | 0.70 | - | - | - |
| MLP | - | - | - | - | - | - | **0.62** | 0.63 | 0.62 |

# Discussion and conclusion:

Overall, our results indicate that it is not necessarily better to incorporate visual elements in text classification with automated analysis. DL based classifier is outperformed by SL models with much simpler data representation by a considerable margin. This result begets insights for automated content analysis. While visual materials can be important for the delivery of the message, they may create more noise than signal in automated content analysis. In our case, the natural public relations visual content was extremely diverse often including creative visual infographics. These images are, by nature, are rather different from training images often used in computer vision(CV) models. Moreover, the substantive purposes of computer vision and social sciences often diverge from each other. Therefore, CV models tend to have a different purpose. This complicates choosing the right CV model even further. These circumstances clearly calls for visual material processing models tailored for social science purposes. While this would be a rather demanding task, it should be possible to create such models using existing machine learning architectures.

Based on our results, there are several good practices we can recommend for the future. First of all, DL algorithms require large amount of data to reach acceptable performance. Therefore, it is always wise to start simple. As our results show, SL algorithms with simpler featurization can accomplish the task even if they do not encode information from visual materials. Therefore, our first recommendation is to test the simpler alternative. However, if the research question inextricably requires a multimodal analysis, we recommend two key actions. First of all, it is best to use DL algorithms with multi-modal embeddings as they can handle high-dimension tensors better than SL algorithms. However, for this we recommend researchers to have more than 1000 labelled observations for their model. Finally, for those researchers with limited resourece, we would like to point out that it is still possible to use SL algorithms with multi-modal embeddings. However, SL algorithms are not designed to handle tensors. This requires the researcher to apply dimension reduction to multi-modal embeddings which may lead to substantial information loss as our experiments show. Therefore, it is imperative to find the optimum dimension reduction method before employing this option.

# References:

Benoit, K., Watanabe, K., Wang, H., Nulty, P., Obeng, A., Müller, S., & Matsuo, A. (2018). quanteda: An R package for the quantitative analysis of textual data. *Journal of Open Source Software*, *3*(30), 774. https://doi.org/10.21105/joss.00774

Chollet, F. (2015). *Keras*. GitHub. https://github.com/fchollet/keras

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

Kuhn, M. (2008). Building Predictive Models in *R* Using the **caret** Package. *Journal of Statistical Software*, *28*(5). https://doi.org/10.18637/jss.v028.i05

Li, K., Zhang, Y., Li, K., Li, Y., & Fu, Y. (2022). Image-Text Embedding Learning via Visual and Textual Semantic Reasoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *PP*. https://doi.org/10.1109/TPAMI.2022.3148470

Li, L. H., Yatskar, M., Yin, D., Hsieh, C.-J., & Chang, K.-W. (2019). VisualBERT: A Simple and Performant Baseline for Vision and Language. *ArXiv:1908.03557 [Cs]*. http://arxiv.org/abs/1908.03557

Loper, E., & Bird, S. (2002). *NLTK: The Natural Language Toolkit* (arXiv:cs/0205028). arXiv. https://doi.org/10.48550/arXiv.cs/0205028

Niu, Y., Lu, Z., Wen, J.-R., Xiang, T., & Chang, S.-F. (2019). Multi-Modal Multi-Scale Deep Learning for Large-Scale Image Annotation. *IEEE Transactions on Image Processing*, *28*(4), 1720–1731. https://doi.org/10.1109/TIP.2018.2881928

Özdemir, S., & Rauh, C. (2022). A Bird’s Eye View: Supranational EU Actors on Twitter. *Politics and Governance*, *10*(1), 133–145. https://doi.org/10.17645/pag.v10i1.4686

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *The Journal of Machine Learning Research*, *12*(null), 2825–2830.

Tseng, S.-Y., Narayanan, S., & Georgiou, P. (2021). Multimodal Embeddings From Language Models for Emotion Recognition in the Wild. *IEEE Signal Processing Letters*, *28*, 608–612. https://doi.org/10.1109/LSP.2021.3065598

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., … Rush, A. M. (2020). HuggingFace’s Transformers: State-of-the-art Natural Language Processing. *ArXiv:1910.03771 [Cs]*. http://arxiv.org/abs/1910.03771

Wu, Y., Kirillov, A., Massa, F., Lo, W.-Y., & Girshick, R. (2019). *Detectron2*. https://github.com/facebookresearch/detectron2