# 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 grew in volume and complexity. The advances in computational methods, such as dictionary analysis or supervised machine learning, allow researcher to process vast quantities of text. Especially with recent advances in computer science, such methods became more convinient to use with the increased availability and ease-of-use of softwares such as Quanteda (Benoit et al., 2018), NLTK (Loper & Bird, 2002), caRet(Kuhn, 2008) and SciKitLearn(Pedregosa et al., 2011).

While automated content analysis methods offer a chance to analyze large quantites of text, dictionary and shallow learning (SL) based methods are still rather costly. They often require considerable amount of effort from researcher to create training datasets, extensively test and validate their models (Grimmer & Stewart, 2013). Moreover, these methods are often unable to help researchers process multi-modal communication where the message is spread across different modalities such as memes, and campaing posters.

While deep learning (DL) based models have similar, if not more demanding, requirements in terms of training and validation, they also carry possiblity to extend content analysis inquiries to multi-modal materials. Previous studies have demonstrated the flexibility of embeddings to analyze multimodel data(Li et al., 2022; Niu et al., 2019; Tseng et al., 2021). This potential becomes even more valuable as the political and social actors adapt and increase their communication on social media platforms using audio-visual materials as well as text. 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 (Özdemir & Rauh, 2022). Against this backdrop, DL embeddings carry the potential to open up new research venues and possiblities.

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 utilize series of SL models and multi-modal DL embedding to classify manually annotated tweets from EU executives. We then compare the classification performance of these models. Our results indicate [...]. 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 830 tweets from the EU executive accounts sent out between December 1st 2019 and July 31st 2020. We focus on these 828 tweets in our experiment specifically because they deliver their message by combining text and images or supplemented the textual message with imagery. An example of them is presented in figure 1. The tweets are manually annotated as a part of the project Trondheim Analytica (Özdemir, Graneng, de Wilde, forthcoming). The orginal dataset is composed of all tweets from the 117 verified accounts of EU executives which includes commissioners, director generals, institutions and agencies responsible for policy making and implementation at the EU level. Tweets were manually coded as a whole according to its key message by a team of two researchers. These key messages, object of publicity, are the specific acts in the tweet such as meetings, identity or mission statements, opinions, and identified by “what is being publicized in the tweet?”. Object of publicity consists of six different categories which are coded as individual binary indicators since tweets can contain multiple categories of object of publicity. Reply category is provided by Twitter API. Categories for object of publicity and example tweets are presented in Table 1. To ensure the data quality, three rounds of intercoder reliability tests were conducted between coders before coding the full sample. The first two rounds showed insufficient reliability scores, but after intensive training and discussion among coders, we reached sufficient reliability scores across all coding categories (Krippendorf α >.8).

Table 1: Categories for object of publicity

|  |  |  |  |
| --- | --- | --- | --- |
| **Object of publicity categories** | **Definition** | **Label distributions**  **(N=828)** | |
| **1** | **0** |
| Identity and mandate | Messages that aim to inform the audience about reasons as to why the EU, its institutions and bureaucrats exist and have a political role | 72 | 756 |
|  |  |  |  |
| Output | Messages that provide an update and information on political operations, policies, programs, reports published by the EU, its institutions, or its bureaucrats | 449 | 379 |
|  |  |  |  |
| Activity | Tweets containing activities such as: meetings, handshakes, travel, signing documents, conference participation by officials that show actions or events taking place outside of Twitter. This could be actions by the account holder or others. | 302 | 526 |
|  |  |  |  |
| Opinion | Tweets that state the author’s preference or evaluation regarding some policy, activity, situation or institutions and actors. | 194 | 634 |
|  |  |  |  |
| Other | Tweets that do not pertain to political or day to day operations defined in author’s mandate such as job announcements or trivia information. | 68 | 760 |
|  |  |  |  |
| Input seeking | Tweets that seek feedback, input, or opinions of stakeholders or the wider audience on political operations of the EU. | 25 | 803 |

Graphical user interface, text

Description automatically generated

Figure 1: Example tweets

In our experiment, we take each category as a separate binary outcome and learn individual SL and DL predictive models for them. From SL classification family we use Naïve Bayes (NB), Support Vector Machine (SVM), Random Forest (RF), logistic regression and XGBoost (XGB) while we use XXXX from DL classification family. We extensively use Quanteda (Benoit et al., 2018) and caRet (Kuhn, 2008) packages in R for text preprocessing and SL models. For DL models, we utilize XXXX in python.

In our experimental setting, we test the predictive capacity of these models on three different datasets. First dataset contains only the preprocessed textual part of sampled tweets. The second dataset contains the only the preprocessed images. The third dataset combines textual and visual elements as input data. We, then, learned separate models for binary outcomes for six categories. Finally, we compare the predictive capacity of all models across three different datasets using F1 scores.

# Results:

Table 2: F1 scores of predictive models

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | DFM | | | TFIDF | | | Multi-modal embeddings | | |
|  | F1 | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall |
| Naïve Bayes | 0.70 | 0.54 | 1.00 | 0.70 | 0.54 | 1.00 | - | - | - |
| **Logistic regression** | 0.92 | 0.91 | 0.93 | 0.93 | 0.92 | 0.94 | - | - | - |
| Support Vector machine | 0.69 | 0.75 | 0.64 | 0.69 | 0.68 | 0.69 | - | - | - |
| **Random forest** | 0.94 | 0.94 | 0.95 | 0.93 | 0.95 | 0.92 | - | - | - |
| XGBoost | 0.89 | 0.88 | 0.90 | 0.84 | 0.85 | 0.84 | - | - | - |
| DeepL |  |  |  |  |  |  |  |  |  |