Automatic Emotions Analysis for French Email Campaigns Optimization

Alexis Blandin
IRISA, EXPRESSION,
Université Bretagne-Sud
UNEEK - Kosmopolead
France
email: alexis.blandin@univ-ubs.fr

Farida Said
IRISA, EXPRESSION,
LMBA
Université Bretagne-Sud
France
email: farida.said@univ-ubs.fr

Jeanne Villaneau, Pierre-François Marteau
IRISA, EXPRESSION,
Université Bretagne-Sud
France
email: jeanne.villaneau@univ-ubs.fr,
email: pierre-francois.marteau@univ-ubs.fr

***Abstract*—Email communication and newsletter campaigns remain a significant concern for companies. The main question addressed here is how to optimize the form and content of a newsletter so that it is not interpreted as spam or annoyance by the recipient. We address this question by analyzing the emotions and opinions conveyed by emails and evaluating how they affect their open and click rate performance. We first describe a new dataset of French newsletters, and then we use emotional embeddings to analyze the associations between emotions and email performance. We finally derive clues on how to write effective email campaigns.**

Keywords—Algorithm; Artificial intelligence; Sentiment Analysis; Emotion prediction; Emotion recognition; Email campaign

I. Introduction

Artificial intelligence is developing in many areas and is increasingly used to determine and optimize business and marketing strategies. In particular, Natural Language Processing (NLP) techniques are widely used for the automatic analysis of human interactions, and we exploit them to optimize email communication by analyzing the content of newsletters.

More precisely, we focus on how emotions and opinions conveyed in an emailing campaign can influence its performance. To address this question, we first built a dataset of more than 900 French newsletter campaigns provided by various companies or associations.

We first proposed vector representations of newsletters that reflect emotion and sentiment using NLP techniques. We then statistically analyzed the relationships between the emotions and opinions conveyed by the newsletters and their performance indicators, i.e., click and open rates. Finally, we used the proposed vectorizations to evaluate the prediction of a newsletter's performance based on the emotions and opinions in its text.

II. Related work

A. Marketing studies

We first review some hypotheses formulated in marketing science regarding email marketing optimization and their potential links to emotions. A study proposed in 2008 by K. Byron [1] suggests that the lack of face-to-face interaction due to email communication can lead to misinterpretation of emotions.

According to the author, the lack of cues that allow the recipient to determine the intended emotions generally leads to a *neutrality effect*. The design of the email may even increase the likelihood that the recipient will perceive the email negatively, resulting in a *negativity effect*. The author argues that when the email contains few cues about emotions, the ambiguity of the emotional tone increases the salience of all negative information. For instance, sarcasm may be perceived more negatively than in face-to-face interaction because of the lack of context and tone ambiguity. The study also highlights the importance of the social context of email communication and the socio-demographic characteristics of the sender and recipient (gender, age, relative status in the company, Etc.) in interpreting emotions. Although the author points out some positive consequences of the negativity effect, such as "*using less niceties or not "sugarcoating" the message*," it should be noted that the negativity effect can be problematic in the context of marketing communication in which it is crucial to elicit positive emotions such as pride [2] in order to expect better actions from the customer, especially in western culture.

Furthermore, when the sender and recipient do not know each other, there is even less contextual information to help the recipient interpret the emotions correctly. Thus, there is an increased risk of misinterpreting the emotions conveyed by the email. On the other hand, in a recent study conducted in the French context of the COVID-19 pandemic [3], commercial communication by email seems to evolve from purely informative content to more entertaining and emotional content. Therefore, it is becoming crucial for companies that the recipients do not misinterpret the emotions contained in their emails.

B. Email content analysis

As mentioned before, we want to evaluate the impact of textual content and email subject lines on the performance of email campaigns by analyzing the emotions and opinions they convey. This approach was used in [4] with email subject lines, and their findings validated some hypotheses on how emotions influence email perception. However, these results are difficult to transfer to our context for several reasons, including language. Indeed, the authors investigated the Enron dataset [5], a large set of emails in English from 150 employees, mainly executives, of the Enron company. This type of resource does not seem to exist in French, and many non-English speaking studies have to build their datasets specifically for their tasks [6], [7].

 ISBN: 978-1-61208-896-9

Another major difference with the cited study is that we consider both the subject and the textual content of the email.

C. Emotion Detection

In recent years, emotion detection in text has become increasingly popular due to its wide range of applications. It can be viewed as an extension to a more diverse emotional spectrum of research on sentiment analysis which focuses on positive and negative emotions. While many studies have proposed their own approaches [8], one of the most common is to use word lexicons labeled with categories of emotions. These categories are often the six basic emotions proposed by P. Ekman [9]: joy, fear, disgust, sadness, anger, and surprise.

This type of resource exists in French. A. Abadoui et al. proposed the FEEL lexicon [10], composed of French words or expressions represented by zero-one vectors of size 7. Six entries indicate whether the word carries one of the six basic emotions, and one represents the polarity associated with the word. This lexicon was constructed automatically, from the English NRC-Emolex lexicon, by crossing the results of several automatic translators. A professional translator subsequently enriched the lexicon and validated the results. The final lexicon consists of 14,127 different lemmatized terms, including 11,979 simple words and 2,148 compound words. Each lemmatized form gets the emotions contained in all of its inflected forms.

III. Dataset presentation

Our dataset is composed of newsletters from various organizations such as companies and associations. These organizations use the same customer relationship management (CRM) system and design their emailing campaigns using the same framework. The main objective of these organizations is to inform their subscribers about events or new opportunities. Our dataset does not include campaigns that target purchase actions such as online shopping.

A newsletter's performance can be measured by tracking the included links, which provide the number of unique opens and the number of unique clicks generated by the reader for each newsletter. These are good indicators of the performance of an email campaign and are commonly used in email analysis [11], [12]. One can view the open rate as a measure of the email's attractiveness and the click rate as the engagement generated by the newsletter.

After cleaning up the data provided by the CRM servers and removing test emails and duplicates, we ended up with 973 newsletters, each sent to multiple subscribers, with their performance information such as click rates and open rates. The number of emails per customer is not balanced, as illustrated in Figure 1. While this could represent a bias, we assume it does not impact our analysis. Indeed, we focus on features that can be considered independent of the email's author.

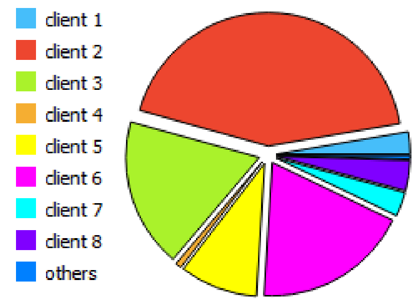

Fig. 1. Distribution of newsletters per client

IV. Data analysis

A. Features extraction

To process the data, we collected descriptive, emotional and sentiment information about the newsletters. The descriptive features were obtained directly from the data host and consist of email subject line length, size of the *.eml* file sent, and unique open and click rates.

We used standard NLP techniques to assess the emotion and sentiment features of the newsletters. First, we segmented the textual content of each newsletter into sentences from which we extracted all words, excluding the French stop words.

For emotion analysis, we assigned to each word an emotion vector according to the FEEL lexicon [10]. If a word is an inflected form, we consider the vector associated with the lemma as the aggregation of all emotions contained in its inflected forms. Then, we computed the emotion vector of each sentence as the average of the emotion vectors of its constituent words. These vectors, constructed from the FEEL lexicon, represent the six basic emotions described by P. Ekman

For sentiment analysis, we evaluated the subjectivity and polarity of the newsletters using the free NLP tool *Python TextBlob for Natural Language Processing*. The textblob library, detailled by Klein and Loper [13], uses a built-in model to compute subjectivity and polarity scores of sentences. A subjectivity value close to 0 indicates objective text, while a value close to 1 indicates highly subjective text. Polarity values range from -1 to 1 and reflect the negativity or positivity of a sentence, respectively.

At this point, each sentence is represented by a vector with eight entries: six emotion scores from the FEEL lexicon and two scores from the textblob analysis. For the full content of the email, we aggregated the information from all sentences by taking the average of all sentence vectors.

In addition to analyzing the newsletter content, we were interested in the emotions conveyed by the email subject line, its polarity, and subjectivity. We, therefore, performed the same NLP processing as for the textual content of the email by considering the subject line as a single sentence. We observed that the emotion scores were almost all null, which led us to consider only the polarity and subjectivity.

In the end, we represent the newsletters by vectors with ten entries that express the emotions and opinions conveyed by the content and topic of the email. Other studies [11] have combined some of the features we consider with non-emotional features, and we question whether emotion and sentiment features are as discriminative as non-emotional features in predicting performance.

B. Statistical results

We explored in our dataset the relationships between the emotion and sentiment features of the newsletters and their performance indicators, namely open and click rates.

TABLE I
PEARSON CORRELATIONS BETWEEN THE CHARACTERISTICS OF THE NEWSLETTERS AND THEIR PERFORMANCE INDICATORS

Features	Open rate	Click rate
File size (FS)	-0.14***	0.25***
Subject line length (SL)	-0.13***	0.18***
Subject line polarity (SP)	-0.07**	-0.03^{ns}
Subject line subjectivity (SS)	-0.01^{ns}	-0.07*
Content Polarity (CP)	-	0.09**
Content Subjectivity (CS)	-	-0.07*
Content Joy (J)	-	-0.10**
Content Fear (F)	-	-0.11***
Content Sadness (S)	-	-0.23***
Content Anger (A)	-	0.06^{ns}
Content Surprise (Su)	-	-0.11***
Content Disgust (D)	-	-0.07*

*p-value < .05, **p-value < .01, ***p-value < .001,^{ns} not significant

The results are presented in Table I in terms of Pearson correlation. It appears that classical descriptors such as subject line length or file size significantly correlate with performance. Indeed, longer subject lines or heavier emails are associated with fewer opens but more clicks if the email is opened.

More interestingly, all emotions conveyed by the email content are negatively correlated with the click rate, regardless of the type of emotion. Sadness is the emotion most negatively associated with the click rate: the more sad the content of the email, the fewer clicks are measured. On the other hand, Table II sheds light on the relationships between the features of the newsletters. One can see that polarity and subjectivity are positively associated both in the text's content and in the email's subject line. Content polarity is positively associated with all emotions except fear and disgust. The significance of the correlations is even greater between content subjectivity and emotions except for disgust. Finally, subject line subjectivity is positively associated with all emotions except disgust, while its polarity is only associated with joy.

On the other hand, emotions are, for the most part, positively associated with each other. If we focus on the highly significant correlations, in bold in the table, we can see that surprise is positively correlated with all emotions except joy and that disgust is associated with rather negative emotions (fear, sadness, anger, and surprise). It also appears that fear and sadness are particular emotions by their strong association and their high correlation with all the emotions.

TABLE II
PEARSON CORRELATIONS BETWEEN THE FEATURES OF THE NEWSLETTERS

	SP	SS	CP	CS	J	F	Sa	A	Su	D
SP	1	0.49***	0.06^{ns}	-0.02^{ns}	0.14***	0.02^{ns}	0.06^{ns}	-0.03^{ns}	0.04^{ns}	0.02^{ns}
SS	-	1	0.04*	0.07^{ns}	0.12***	0.14***	0.11***	-0.02^{ns}	0.15***	0.07*
CP	-	-	1	0.4***	0.1**	0.02^{ns}	0.12***	0.1**	0.2***	0.02^{ns}
CS	-	-	-	1	0.1**	0.14***	0.21***	0.12***	0.2***	-0.03^{ns}
J	-	-	-	-	1	0.19***	0.14***	0.01^{ns}	0.04^{ns}	0.07*
F	-	-	-	-	-	1	0.61***	0.37***	0.32***	0.37***
Sa	-	-	-	-	-	-	1	0.21***	0.25***	0.34***
A	-	-	-	-	-	-	-	1	0.08**	0.27***
Su	-	-	-	-	-	-	-	-	1	0.17***
D	-	-	-	-	-	-	-	-	-	1

*p-value < .05, **p-value < .01, ***p-value < .001,^{ns} not significant
SP: subject line polarity, SS: subject line subjectivity, CP: content polarity, CS: content subjectivity
J: joy, F: fear, Sa: sadness, A: anger, SU: surprise, D: Disgust

Emotion and sentiment features may not be the best predictors of newsletter performance, but we propose to evaluate their effectiveness in predicting click rate in the following.

V. UNSUPERVISED CLUSTERING

A. Multidimensional representation

We first explored our data graphically to see if there is a global structure that we could exploit. Since the newsletters are represented in a 10-dimensional space, we used a dimensionality reduction technique, namely the t-SNE (t-distributed Stochastic Neighbor Embedding). This method is mainly used to project high-dimensional data into low-dimensional spaces (2D or 3D) while preserving local distances between data points.

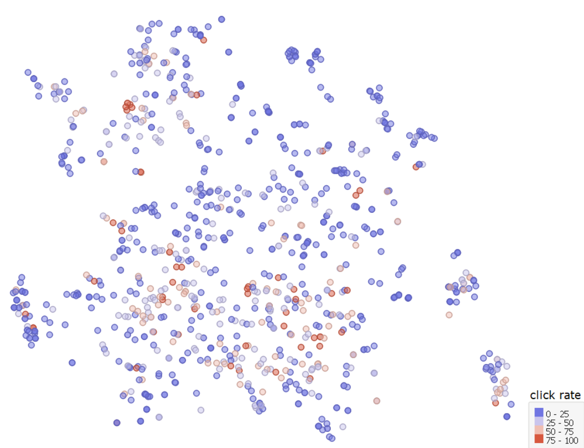

Fig. 2. t-SNE projection of our dataset

Figure 2 gives a visualization of our data in a 2D map. The color associated with the newsletters ranges from blue for "low-performing or bad" newsletters to red for "high-performing or good" newsletters. There is no clear separation between "good" and "bad" newsletters. However, it appears that the "good" newsletters are more grouped while the "bad" ones are more scattered across the map.

We investigate this hypothesis in the next section, using unsupervised clustering.

B. K-Means approach

K-means clustering is a vector quantization method that aims at partitioning a dataset into k clusters by assigning each observation to the cluster with the closest center (or centroid).

The effectiveness of K-means depends highly on the chosen number of clusters, and we usually do not have prior knowledge of the number of clusters that correspond to the most relevant clustering. A measure of the effectiveness of a cluster is the silhouette coefficient as introduced by Kaufman and Rousseeuw [14]. It measures how similar an object is to its cluster (cohesion) compared to other clusters (separation). The value of the silhouette ranges between -1 and 1, where a high value indicates that the object is well matched to its cluster and poorly matched to neighboring clusters. We compute the silhouette coefficients of all points and average them to obtain a global silhouette score. The clustering configuration with the best global silhouette score is the most relevant.

For a clustering in k clusters, the cohesion of a data point i assigned to a cluster I_k is defined as:

$$a(i) = \frac{1}{|I_k| - 1} \sum_{j \in I_k, j \neq i} d(x^i, x^j) \tag{1}$$

where $d(x^i, x^j)$ stands for the distance between the representative vectors x^i and x^j. We chose the cosine distance based on the angle between vectors for its efficiency in clustering textual data. The separation of point i is its average distance to all points in the closest cluster to its cluster I_k:

$$b(i) = \min_{k' \neq k} \frac{1}{|I_{k'}| - 1} \sum_{i' \in I_{k'}} d(x^i, x^{i'}) \tag{2}$$

The silhouette coefficient of point i is then computed as:

$$\begin{aligned} s_{silhouette}(i) &= \frac{b(i) - a(i)}{\max\left(a(i), b(i)\right)} \quad & if : |C_i| > 1, \\ s_{silhouette}(i) &= 0 & if : |C_i| = 1 \end{aligned} \tag{3}$$

We aim to cluster the newsletters using only their emotion and sentiment features. However, some of these features are significantly correlated, as shown in Table II. We then used PCA to denoise the data and construct a representation free of redundant information.

We, therefore, had two hyperparameters to determine: the appropriate number of principal components and the optimal number of clusters. To this end, we used two criteria: the ratio of variance explained by the PCA components and the clustering silhouette score. Table III gives, for each number of principal components, its explained variance rate, its associated optimal clustering and the corresponding silhouette score.

It appears that the partition into two clusters is the best clustering configuration for most of the PCA representations.

We decided to consider eight principal components in the subsequent analysis because they account for more than 91% of the variance in our data set.

TABLE III
OPTIMAL CLUSTERINGS ASSOCIATED WITH DIFFERENT REPRESENTATIONS OF THE DATA

PCA[a]	Explained variance	Number of clusters[b]	silhouette score
1	24%	2	0.577
2	40%	2	0.501
3	53%	4	0.411
4	63%	2	0.358
5	72%	2	0.274
6	79%	2	0.269
7	86%	3	0.250
8	91%	2	0.258
9	96%	4	0.392
10	100%	4	0.366

[a] Number of PCA components
[b] The optimal number of clusters is chosen to maximize the silhouette score

We compared the performance of the two resulting clusters in terms of click rates, but the results were not conclusive. The distributions of the click-through rates in the two clusters are presented in Figure 3.

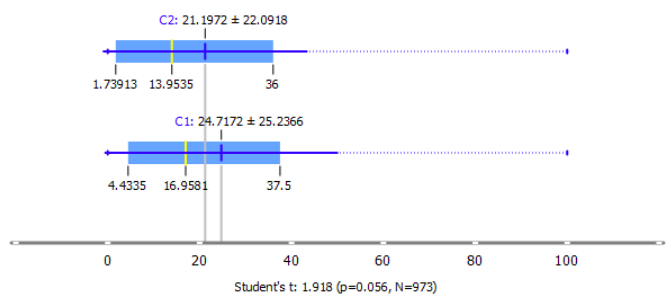

Fig. 3. Click rates in the two clusters obtained by k-means with 8 principal components

As a result, the emotion and sentiment features alone did not allow us to discriminate between "good" and "bad" newsletters. Nevertheless, in the next section, we use another approach based on supervised classification.

VI. CLASSES OF PERFORMANCE PREDICTION

A. Classes of performance

To implement supervised classification, we need to label our data set. Following the considerations in the previous section, we decided to create two performance classes around the median click rate. One class contains the 50% of newsletters that generate the fewest clicks, and the other class contains the highest click rates. We refer to them as the "poor or lower-performing" class and the "good or higher-performing" class.

Figure 4 gives the distribution of data silhouette scores by performance class. Here, the embeddings cover all emotion and sentiment features.

We observe that newsletters with lower click rates are more dispersed around their class center than better performing newsletters. This trend is even more marked when we do not

 ISBN: 978-1-61208-896-9

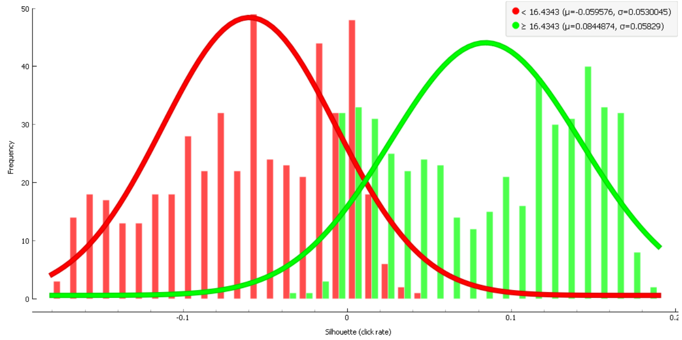

Fig. 4. Distribution of silhouette scores in the "bad" class (red) and the "good" class (green), with subject line features

take into account the subject line's subjectivity and polarity, as shown in Figure 5.

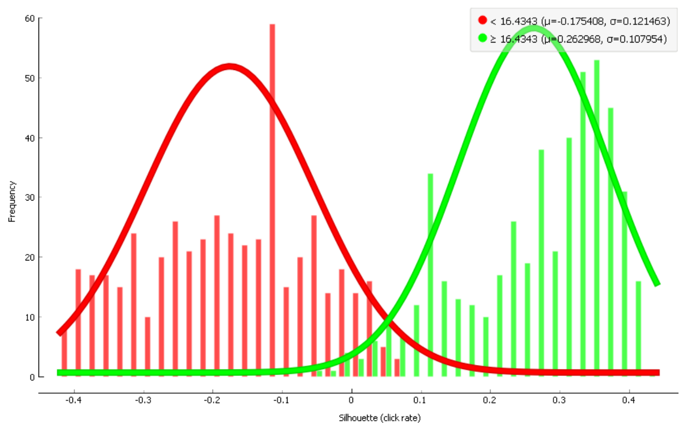

Fig. 5. Distribution of silhouette scores in the "bad" class (red) and the "good" class (green), without subject line features

B. *Supervised classification*

We are interested in evaluating the predictive power of the ten emotion and sentiment features of the newsletters, categorized into bad and good newsletters as defined above.

For this purpose, we performed classifications of our dataset with various machine learning methods [15]. Table IV presents their performance measures estimated with 10-fold cross-validation procedure.

It appears that the best classifiers are AdaBoost, Neural Network, and Random Forest. We also notice that the performance scores are very slightly lower without the subject line information.

To measure the contribution of each feature in the predictive model, we tested "leave-one-out" and "one-at-once" procedures with the best classifier, Adaboost. In leave-one-out experiments, we considered all but one feature, and in one-at-once experiments, we considered one feature at a time. The F1-scores presented in Table V are to be compared with the F1-score of the full model, constructed with all predictors, which is 0.723 (see Table III).

TABLE IV
PERFORMANCE SCORES OF THE CLASSIFIERS, WITH AND WITHOUT SUBJECT LINE INFORMATION

Classifier	F1 Score	Precision	Recall
With subject line information			
AdaBoost	**0.723**	**0.724**	**0.724**
Neural Network	0.712	0.712	0.712
Random Forest	0.711	0.711	0.711
kNN	0.681	0.688	0.683
Naive Bayes	0.666	0.666	0.666
SVM	0.607	0.617	0.612
Logistic Regression	0.585	0.594	0.590
Constant	0.500	0.500	0.500
Without subject line information			
Model	**F1 Score**	**Precision**	**Recall**
AdaBoost	**0.722**	**0.723**	**0.723**
Neural Network	0.714	0.715	0.715
Random Forest	0.710	0.710	0.710
kNN	0.679	0.683	0.680
Naive Bayes	0.666	0.666	0.666
SVM	0.628	0.640	0.633
Logistic Regression	0.621	0.643	0.630
Constant	0.500	0.500	0.500

TABLE V
ADABOOST PERFORMANCE SCORES WITH A SINGLE FEATURE OR ALL BUT ONE FEATURE

Feature	F1-score with a single feature	F1-score with all but one feature
Subject line polarity	0.498	0.720
Subject line subjectivity	0.503	0.721
Content Polarity	0.614	0.719
Content Subjectivity	0.570	0.725
Content Joy	0.624	0.723
Content Fear	0.604	0.722
Content Sadness	**0.633**	**0.711**
Content Anger	0.618	0.713
Content Surprise	0.614	0.721
Content Disgust	0.626	0.721

These results confirm the impact of emotions and sentiment on newsletter click rates, and as observed in Section IV, sadness is the emotion with the most impact. We can also see that text content subjectivity has a negative effect on prediction. We should investigate these observations further to improve our embeddings in future work.

VII. CONCLUSION

In this paper, we explored to what extent emotion and sentiment detection can help predict the performance of an email campaign. Literature in the marketing field suggests that email communication generally results in a misunderstanding of the emotions being conveyed and, due to negativity and neutrality effects, these emotions are often misinterpreted as neutral or negative by the recipient. When the recipient is a potential customer or subscriber, this negative effect can lead to unwanted behavior, measured with objective metrics such as open rate or click rate.

We presented a dataset composed of French emailing campaigns and represented them with emotion and sentiment

 ISBN: 978-1-61208-896-9

embeddings. Our study shows that almost all emotions are negatively correlated with newsletter performance, especially sadness.These results are consistent with the marketing literature, which suggests that negative emotions, such as sadness, are well identified by the recipient, while positive emotions or opinions in a text (represented by the subjectivity score) are poorly understood.

In addition, we observed that the best-performing newsletters have more homogeneous emotion and sentiment features than the less-performing newsletters. This finding needs further investigation to build a guide for writing effective newsletter.

Finally, we used emotion and sentiment embeddings to predict performance classes of our newsletters. The presented approach is perfectible, but it already constitutes a good baseline for our future work on emotion detection in French emails. Areas of improvement concern, in particular, the hyper-parameters of the classifiers and the embeddings. Moreover, we will soon provide the scientific community with our dataset to enrich the French resources and allow interested researchers to reproduce and improve our work.

REFERENCES

[1] K. Byron, "Carrying too heavy a load? the communication and miscommunication of emotion by email," *The Academy of Management Review*, vol. 33, no. 2, pp. 309–327, 2008, Accessed on Aug. 18, 2021. [Online]. Available: http://www.jstor.org/stable/20159399

[2] J.-E. Kim and K. Johnson, "The Impact of Moral Emotions on Cause-Related Marketing Campaigns: A Cross-Cultural Examination," *Journal of Business Ethics*, vol. 112, no. 1, pp. 79–90, January 2013, Accessed on Aug. 18, 2021. [Online]. Available: https://ideas.repec.org/a/kap/jbuset/v112y2013i1p79-90.html

[3] M. S.-F. Virginie Rodriguez, "Le contenu des communications relationnelles par email des enseignes : Quelle perception par le consommateur ? [Content of retailers' relational e-mails: what is the consumer's perception ?]," in *20th International Marketing Trends Conference*, Venise, Italy, Jan. 2021, Accessed on Aug. 29, 2021.

[4] R. Miller and E. Charles, "A psychological based analysis of marketing email subject lines," in *2016 Sixteenth International Conference on Advances in ICT for Emerging Regions (ICTer)*, 2016, pp. 58–65.

[5] B. Klimt and Y. Yang, "The enron corpus: A new dataset for email classification research," in *Machine Learning: ECML 2004*, J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004, pp. 217–226.

[6] R. Kalitvianski, "Traitements formels et sémantiques des échanges et des documents textuels liés à des activités collaboratives[Formal and semantic processing of exchanges and textual documents related to collaborative activities.]," Theses, Université Grenoble Alpes, Mar. 2018, Accessed on Sep. 3, 2021. [Online]. Available: https://tel.archives-ouvertes.fr/tel-01893348

[7] H. Guenoune, K. Cousot, M. Lafourcade, M. Mekaoui, and C. Lopez, "A dataset for anaphora analysis in French emails," in *Proceedings of the Third Workshop on Computational Models of Reference, Anaphora and Coreference*. Barcelona, Spain (online): Association for Computational Linguistics, Dec. 2020, pp. 165–175, Accessed on Aug. 27, 2021. [Online]. Available: https://aclanthology.org/2020.crac-1.17

[8] A. Seyeditabari, N. Tabari, and W. Zadrozny, "Emotion Detection in Text: a Review," p. arXiv:1806.00674, Jun. 2018, Accessed on Sep. 1, 2021.

[9] P. Ekman, *Basic Emotions*. John Wiley & Sons, Ltd, 1999, ch. 3, pp. 45–60, Accessed on Aug. 21, 2021. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/0470013494.ch3

[10] A. Abdaoui, J. Azé, S. Bringay, and P. Poncelet, "FEEL: a French Expanded Emotion Lexicon," *Language Resources and Evaluation*, vol. 51, no. 3, pp. 833–855, Sep. 2017, Accessed on Sep. 3, 2021. [Online]. Available: https://hal-lirmm.ccsd.cnrs.fr/lirmm-01348016

[11] A. Kumar, "An empirical examination of the effects of design elements of email newsletters on consumers' email responses and their purchase," *Journal of Retailing and Consumer Services*, vol. 58, p. 102349, 2021, Accessed on Aug. 29, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0969698920313576

[12] A. Bonfrer and X. Drèze, "Real-time evaluation of e-mail campaign performance," *Marketing Science*, vol. 28, no. 2, p. 251–263, 2009, Accessed on Aug. 22, 2021. [Online]. Available: https://doi.org/10.1287/mksc.1080.0393

[13] U. Yaqub, S. A. Chun, V. Atluri, and J. Vaidya, "Analysis of political discourse on twitter in the context of the 2016 us presidential elections," *Government Information Quarterly*, vol. 34, no. 4, pp. 613–626, 2017, Accessed on Aug. 31, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/ S0740624X17301910

[14] L. Kaufman and P. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons., 1990, john Wiley & Sons, New York.

[15] A. Mueller and S. Guido, *Machine learning avec Python*. O'Reilly Media, Inc., 2018.

 ISBN: 978-1-61208-896-9