Multimodal intent classification with incomplete modalities using text embedding propagation

Victor Machado Gonzaga*
machado.prx@usp.br
Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP)
São Carlos, São Paulo, Brazil

Nils Murrugarra-Llerena*
nmurrugarraller@snap.com
Snap Research
S. Monica, California, USA

Ricardo Marcacini*
ricardo.marcacini@icmc.usp.br
Institute of Mathematics and Computer Sciences (ICMC)
University of São Paulo (USP)
São Carlos , São Paulo, Brazil

ABSTRACT

Determining the author's intent in a social media post is a challenging multimodal task and requires identifying complex relationships between image and text in the post. For example, the post image can represent an object, person, product, or company, while the text can be an ironic message about the image content. Similarly, a text can be a news headline, while the image represents a provocation, meme, or satire about the news. Existing approaches propose intent classification techniques combining both modalities. However, some posts may have missing textual annotations. Hence, we investigate a graph-based approach that propagates available text embedding data from complete multimodal posts to incomplete ones. This paper presents a text embedding propagation method, which transfers embeddings from BERT neural language models to image-only posts (i.e., posts with incomplete modality) considering the topology of a graph constructed from both visual and textual modalities available during the training step. By using this inference approach, our method provides competitive results when textual modality is available at different completeness levels, even compared to reference methods that require complete modalities.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Information systems** → **Web mining**.

KEYWORDS

social networks, multimodal learning, network embedding

ACM Reference Format:
Victor Machado Gonzaga, Nils Murrugarra-Llerena, and Ricardo Marcacini. 2021. Multimodal intent classification with incomplete modalities using text embedding propagation. In *Brazilian Symposium on Multimedia and the Web (WebMedia '21), November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3470482.3479636

1 INTRODUCTION

We are experiencing a transformation from text-based to image-sharing communication, where image-based social networks have become very popular for users to express opinions [4]. In this scenario, a relevant task is determining author's intention when publishing a certain content, in which the intent labels are useful for improving recommendation systems and event analysis for social networks [17, 24]. For example, Figure 1 shows two documents (posts) classified as "Advocative" and "Promotive", respectively. Other labels usually explored in intent classification tasks are "Exhibitionist", "Expressive", "Informative", "Entertainment", and "Provocative". Although images are the central information of these social platforms' interaction, a significant part of the posts also have associated textual data, such as captions, tags, and comments. Posts composed of texts and images are a specific type of multimodal documents and represent a challenging scenario for machine learning tasks [3].

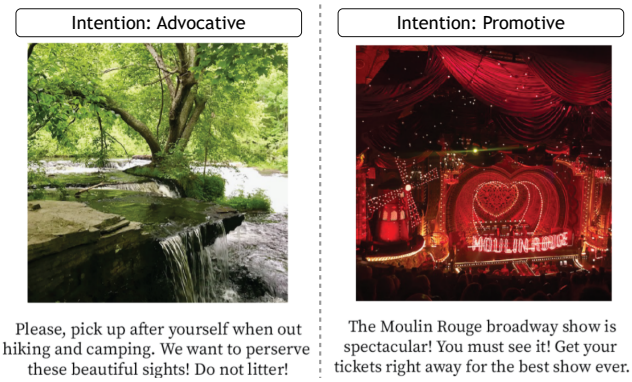

Figure 1: Examples of multimodal document intent (Adapted from [17]).

Classification methods for multimodal documents assume that each modality provides complementary information and the use of different modalities will lead to performance improvements compared to methods based on only one modality [9]. Existing methods explore strategies based on identifying shared structures between modalities [7, 13, 28] and learning abstract representations from modalities using deep neural networks [9]. However, such strategies have limitations in the presence of incomplete modalities, i.e. when a subset of posts do not have textual information. A trivial solution uses only documents with complete modalities, unfortunately discarding data that may contain valuable information.

In this paper, we present a method of text embedding propagation for intent classification in multimodal documents to deal with incomplete textual modality. Determining an author's intention

*Both authors contributed equally to this research.

WebMedia '21, November 5–12, 2021, Belo Horizonte / Minas Gerais, Brazil

ACM ISBN 978-1-4503-8609-8/21/11…$15.00
https://doi.org/10.1145/3470482.3479636

in image-based social network posts requires identifying complex relationships between image and text. Although images are present in all posts, the associated textual information is incomplete (missing modality) or contains ambiguous text, as well as irony and sarcasm. Thus, our method not only deals with the challenge of incomplete modalities but also learns semantic textual representation for all posts through embedding propagation. An overview of the proposed method and our main contributions include:

- We propose a graph-based representation to model multimodal documents. Each vertice represents a multimodal document and can have two associated information vectors: (1) image embeddings obtained by a pre-trained deep residual network [11] and (2) text embeddings obtained by a pre-trained BERT language model [5]. First, documents with complete modalities are used for fine-tuning the embedding models. Second, we generate edges considering the both visual similarity of the images and textual similarity of the captions; and
- We propose a transductive graph learning method to propagate text embeddings from the vertices that contain such information. The multimodal graph structure helps to learn text embeddings for vertices that have only image data. Thus, a node that has only visual features can receive textual features through a process that iteratively propagates BERT embeddings data between neighboring nodes. After embeddings propagation, all vertices will have complete modalities and then any method of multimodal document classification can be used.

We carried out an experimental evaluation involving author intent classification in a multimodal social media dataset. We compared our proposed method with two other methods: (1) an intent document classifier using only the post images; and (2) a state-of-the-art method that combines images and texts for intent classification with complete modalities. We obtained competitive results, even in the presence of incomplete modalities.

2 RELATED WORK

In the last decades, several works on intent classification have been proposed to understand queries from search engine logs, where the intent is usually informational, transactional, or navigational [19, 26]. Recent social media platforms and virtual agents provided new challenges and applications, such as determining the Instagram author's intent [17], intent recognition in doctor-patient interview [25], and intent classification of short-text on Twitter [24].

Given the importance of this intent classification, researchers innovate ways to improve their performance. Larson et al. [18] argues that out-of-scope intents should be included for more robust intent prediction across different domains. Complementary, Gupta et al. [10] propose a joint framework to learn jointly intent and name entity recognition systems. They achieved state-of-the-art results by combining these components. Also, Wallace et al. [27] shows that context is helpful to recognize hard intents such as irony. Our work complements contextual efforts via a multimodal approach, where the textual modality complements the visual modality.

Researchers focus on popular multimodal tasks such as image captioning [1, 6, 16] and visual question answering [2, 8, 12]. These tasks consider textual and visual modalities. Similarly, these modalities are important for ads understanding [13, 28], multimodal classification [14], image retrieval [20, 22] and personalization [21].

Similarly, social media data present these challenges. Detecting hate speech in memes [7, 15] is challenging because the same image with different texts can be identified as hateful or not. Hence, the interaction between textual and visual features is required. The same scenario happens for multimodal intent classification [17], specifically for irony and humor in the entertainment category. Kruk et al. [17] project both complete modalities to the same space vector to learn a classification model, while Gomez et al. [7] employs spatial concatenation and textual kernel models. In contrast, our work tackles these challenges via a graph-based multimodal regularization to learn a data representations in the presence of incomplete modality.

3 PROPOSED METHOD

We investigated the most common scenario involving incomplete modalities for intention classification, in which a subset of multimodal documents (e.g. posts on social networks) contains both visual and textual features. This subset of documents with complete modalities is defined as $\{X^I \in \mathbb{R}^{n \times d^I}, X^T \in \mathbb{R}^{n \times d^T}\}$, where X^I indicates the set of n documents in the d^I-dimensional vector-space of visual features, and X^T indicates the set of n examples in the d^T-dimensional vector-space of textual features. We define $X^I_{inc} \in \mathbb{R}^{m \times d_I}$ as the subset of m documents with incomplete modalities (i.e. missing textual data)

We propose a graph-based representation to identify the relationships between documents in their different modalities. Let $G(V, E, W)$ be a graph, where V is a set of vertices, E a set of edges, and W represents a set of edge weights. Each multimodal document x_i is mapped to a vertex $v_i \in V$. Edges are generated considering two rules as follows:

Rule 1. *Documents with textual features are linked through their nearest k-neighbors from* $X^T \in \mathbb{R}^{n \times d^T}$.
We argue that BERT semantic representations are more effective than visual features for intent classification. Our graph-based representation prioritizes the relationships of textual similarity between documents when this modality is present.

Rule 2. *Documents with visual features are inserted in the graph using the nearest k-neighbors from* $\{X^I \cup X^I_{inc}\} \in \mathbb{R}^{n \times d^I}$.
Visual features are used to generate edges between all documents in the data set. Note that even documents with complete modalities can be reached by this rule, thereby generating paths in the graph that allow the propagation of text embeddings between the vertices.

Our method uses a Gaussian kernel as a non-linear function of Euclidean distance to calculate the graph weights W and thus obtains neighborhood relationships between documents. Given two vertices v_i and v_j, the weight W_{ij} is calculated according to Eq. 1, where x_i and x_j are the respective feature vectors from a vector-space model Φ and σ is a scale parameter. In Rule 1, the vector-space model Φ is composed of the BERT semantic representations. In Rule 2, Φ represents the vector-space model composed of the ResNet visual features. A normalized affinity matrix S is calculated from W,

as defined in Eq. 2. D is a diagonal matrix where the (i,i)-element is equal to the sum of the i-th row of W.

$$W_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2_\Phi}{2\sigma^2}\right) \quad (1)$$

$$S = D^{-\frac{1}{2}} W D^{-\frac{1}{2}} \quad (2)$$

Now, we use the generated graph-based representation to propagate text embeddings between documents. Our method is inspired by graph-based transductive learning proposed by [29] for semi-supervised classification tasks. We have extended the method for text embeddings propagation. The method aims to obtain a new textual representation $F \in \mathbb{R}^{q \times d^T}$ for all multimodal documents, with $q = n+m$, considering both the graph structure (represented by affinity matrix S) and the existing text embeddings (represented by documents $x \in X^T$). In other words, the idea is to learn a $Q : V \rightarrow F$ mapping function, where each vertex v_i must be associated with a new textual embedding f_i.

$$Q(F) = \frac{1}{2} \sum_{v_i, v_j \in V} S_{ij} \|f_i - f_j\|^2 + \mu \sum_{x_i \in X^T} \|f_i - x_i\|^2 \quad (3)$$

We minimize the objective function defined in Eq. 3. The first term $S_{ij}\|f_i - f_j\|^2$ indicates that two documents i and j with high affinity S_{ij} in the graph must have similar text embeddings f_i and f_j, respectively. The second term $\mu \sum_{x_i \in X^T} \|f_i - x_i\|^2$ indicates that the subset of documents with textual embeddings (complete modalities) must preserve their representation according to the parameter μ. The higher the μ value, the greater the preservation of the initial embeddings. Thus, low μ values allow a refinement of the BERT embeddings according to the graph structure.

Eq. 3 can be solved in its closed form using convex optimization methods or via an iterative random-walk algorithm, both obtaining similar solutions [30]. In this paper, we use the iterative version, where the F matrix is randomly initialized. In each iteration, we update the matrix F, where each document spread its text embedding to its neighbors until a global stable state (convergence) is achieved.

4 EXPERIMENTAL EVALUATION

This section presents our experimental analysis using a dataset with 1104 public Instagram posts collected by Kruk et al. [17]. Each post is formed by an image-caption pair and was manually labeled in 7 author's intent labels, as shown in Table 1.

Intent Label	# Samples	Description
Provocative	84	directly attack an individual or group
Informative	119	information regarding an event
Advocative	97	advocate for a figure, idea, or movement
Entertainment	310	entertain using art, humor, memes, etc
Exhibitionist	237	create a self-image reflecting the person
Expressive	95	express emotion at an external entity
Promotive	162	promote events, products, organizations, etc

Table 1: Overview of the dataset used in the experimental evaluation (Adapted from Kruk et al. [17]).

We compare our results with these methods:

Method	ACC	AUC
Chance / Random Classifier	28.1	50.0
Img (Baseline) [11]	42.9 (±0.0)	76.0 (±0.5)
Img+Txt-ELMo (100%) [17]	56.7 (±0.0)	85.6 (±1.3)
Img+Txt-BERT (100%)	58.6 (±0.01)	86.4 (±0.01)
Graph+Txt-Propagation (20%) [ours]	44.0 (±0.02)	77.2 (±0.02)
Graph+Txt-Propagation (40%) [ours]	46.9 (±0.02)	79.3 (±0.01)
Graph+Txt-Propagation (60%) [ours]	51.1 (±0.02)	81.9 (±0.01)
Graph+Txt-Propagation (80%) [ours]	54.5 (±0.01)	84.4 (±0.01)

Table 2: Comparison of the intent classification performance (ACC and AUC). Our method (Img+Txt-BERT) considering different presence percentages of textual information is in bold.

- **Img (Baseline)**: a deep convolutional neural network (DCNN) to classify the post's author intent based only on visual features from ResNet-18 Network [11].
- **Img + Txt-ELMo [17]**: a state-of-the-art DCNN method to classify the post's author intent based on the fusion of visual and textual features. The fusion is via a linear projection of the two modalities in the same embedding space on a DCNN layer. Textual features were learned using the ELMo pre-trained character-based contextual embeddings [23]. Kruk et al. [17] argue that ELMo character embeddings are more robust to noise. This method assumes complete modalities for the entire dataset.
- **Img + Txt-BERT**: the method concatenates the ResNet visual features and BERT textual features in a single representation. Next, uses the SVM method to obtain an intent classification model. The method requires complete modalities and was used as a reference (upper bound) for our proposal.

We use the subset of documents with complete modalities to fine-tune the BERT model. We simulate the following levels of incomplete modalities in the test set: 20%, 40%, 60%, and 80%.

Our graph-based representation for (incomplete) multimodal documents was generated with $k = 30$ for kNN and Euclidean distance using the two proposed rules. Furthermore, we use the average distance between the top-7 nearest neighbors to define the scale parameter σ in Eq. 1. In text embedding propagation, we use $\mu = 1.0$ (Eq. 3) to preserve initial BERT embeddings and propagate these embeddings for incomplete documents according to the graph structure. After the text embedding propagation, our method concatenates the textual and visual features to train an SVM classifier (kernel RBF, parameter $C = 1.0$). We emphasize that the proposed embedding propagation method is responsible for estimating the textual modality of the only-images posts by considering the graph topology.

Table 2 presents a comparison of the classification performance using classification accuracy (ACC) and area under the ROC curve (AUC). The metrics were calculated using the training and test sets previously available in the original dataset (5-fold cross-validation). Our method is referred to as **Graph+Txt-Propagation** in Table 2, followed by the percentage of documents with both modalities.

Our proposed method obtains ACC and AUC metrics greater than the baseline (images only), which indicates that considering some level of textual information increases the author intent classification performance, even with very incomplete modalities, as

shown in Graph+Txt-Propagation (20%). When we consider scenarios with 60% and 80% of posts with textual modality, our proposed method achieves competitive results (ACC and AUC metrics) when compared to methods that require complete modalities Img+Txt-BERT (100%) and Img+Txt-Elmo (100%). We consider such scenarios (60% and 80% of posts with textual modality) to be the closest to the real-world applications involving image-based social networks since a significant percentage of texts are discarded because they are too short, not provided or meaningless.

5 CONCLUDING REMARKS

We propose a method to classify authors' intention in social media datasets with incomplete modalities. We showed that our graph-based representation is a promising structure for combining different modalities, in which we generate edges through visual and textual features. We also showed that BERT-based representation is competitive for the textual modality, even in scenarios with a large percentage of missing texts. Moreover, our method allows to propagate initial BERT embeddings considering the graph topology.

Text embedding propagation showed to be useful to represent posts that contain only images (incomplete textual modality). Our embedding propagation should not be seen as a caption generator, but a method for associating semantic embedding content to images according to neighboring posts. In this scenario, we note that propagated embeddings represent general topics that are promising in determining the author's intention.

Directions for future work involve investigating the performance of the method in adjusting the initial embeddings of the BERT model through the topological structure of the graph, i.e., we plan to reduce the preservation factor of the initial embeddings via parameter μ of Eq. 3. The dataset used in experiments, our source code and more experimental results are publicly available at https://github.com/machadoprx/multimodal-intent-classification.

ACKNOWLEDGMENTS

This work was supported by CNPq [process number 426663/2018-7] and FAPESP [process numbers 2019/25010-5 and 2019/07665-4].

REFERENCES

[1] Jyoti Aneja, Harsh Agrawal, Dhruv Batra, and Alexander Schwing. 2019. Sequential Latent Spaces for Modeling the Intention During Diverse Image Captioning. In *International Conference on Computer Vision (ICCV)*. IEEE.

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*. IEEE.

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence (TPAMI)* (2018).

[4] Tae Rang Choi and Yongjun Sung. 2018. Instagram versus Snapchat: Self-expression and privacy concern on social media. *Telematics and Informatics* (2018).

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. ACL.

[6] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2018. VSE++: Improving Visual-Semantic Embeddings with Hard Negatives. In *British Machine Vision Conference 2018 (BMVC)*. BMVA Press.

[7] Raul Gomez, Jaume Gibert, Lluis Gomez, and Dimosthenis Karatzas. 2020. Exploring Hate Speech Detection in Multimodal Publications. In *Winter Conference on Applications of Computer Vision (WACV)*. IEEE.

[8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition*.

[9] Wenzhong Guo, Jianwen Wang, and Shiping Wang. [n.d.]. Deep multimodal representation learning: A survey. *IEEE Access* ([n. d.]).

[10] Arshit Gupta, John Hewitt, and Katrin Kirchhoff. 2019. Simple, Fast, Accurate Intent Classification and Slot Labeling for Goal-Oriented Dialogue Systems. In *Annual SIGdial Meeting on Discourse and Dialogue*. ACL.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Conference on Computer Vision and Pattern Recognition*.

[12] Drew A. Hudson and Christopher D. Manning. 2019. GQA: A New Dataset for Real-World Visual Reasoning and Compositional Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[13] Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic Understanding of Image and Video Advertisements. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[14] Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, and Davide Testuggine. 2019. Supervised Multimodal Bitransformers for Classifying Images and Text. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop*. NeurIPS.

[15] Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes. *CoRR* (2020).

[16] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *International Journal of Computer Vision (IJCV)* (2017).

[17] Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating Text and Image: Determining Multimodal Document Intent in Instagram Posts. In *Conference on Empirical Methods in Natural Language Processing and the International Joint Conference on Natural Language Processing*.

[18] Stefan Larson, Anish Mahendran, Joseph J. Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K. Kummerfeld, Kevin Leach, Michael A. Laurenzano, Lingjia Tang, and Jason Mars. 2019. An Evaluation Dataset for Intent Classification and Out-of-Scope Prediction. In *Conference on Empirical Methods in Natural Language Processing and International Joint Conference on Natural Language Processing*.

[19] Xiao Li, Ye-Yi Wang, and Alex Acero. 2008. Learning query intent from regularized click graphs. In *Annual international ACM SIGIR conference on Research and development in information retrieval*. ACM.

[20] Nils Murrugarra-Llerena and Adriana Kovashka. 2018. Image Retrieval with Mixed Initiative and Multimodal Feedback. In *British Machine Vision Conference (BMVC)*.

[21] Nils Murrugarra-Llerena and Adriana Kovashka. 2019. Cross-Modality Personalization for Retrieval. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[22] Nils Murrugarra-Llerena and Adriana Kovashka. 2021. Image retrieval with mixed initiative and multimodal feedback. *Computer Vision and Image Understanding* 207 (2021), 103204. https://doi.org/10.1016/j.cviu.2021.103204

[23] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. ACL.

[24] Hemant Purohit, Guozhu Dong, Valerie Shalin, Krishnaprasad Thirunarayan, and Amit Sheth. 2015. Intent classification of short-text on social media. In *International conference on smart city/socialcom/sustaincom (smartcity)*. IEEE.

[25] Robin Rojowiec, Benjamin Roth, and Maximilian Fink. 2020. Intent Recognition in Doctor-Patient Interviews. In *Language Resources and Evaluation Conference (LREC)*. European Language Resources Association.

[26] Manos Tsagkias and Roi Blanco. 2012. Language intent models for inferring user browsing behavior. In *International ACM SIGIR conference on Research and development in information retrieval*. ACM.

[27] Byron C. Wallace, Do Kook Choe, Laura Kertz, and Eugene Charniak. 2014. Humans Require Context to Infer Ironic Intent (so Computers Probably do, too). In *Annual Meeting of the Association for Computational Linguistics (ACL)*. ACL.

[28] Keren Ye and Adriana Kovashka. 2018. ADVISE: Symbolism and External Knowledge for Decoding Advertisements. In *European Conference on Computer Vision (ECCV)*.

[29] Dengyong Zhou, Olivier Bousquet, Thomas Lal, Jason Weston, and Bernhard Schölkopf. 2003. Learning with local and global consistency. *Advances in neural information processing systems (NeurIPS)* (2003).

[30] Xiaojin Zhu and Andrew B Goldberg. 2009. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning* (2009).