MULTIMODAL DATA EVALUATION FOR CLASSIFICATION PROBLEMS

Daniela Moctezuma[1]and Víctor Muníz[2]and Jorge García[2]

[1]Centro de Investigación en Ciencias de Información Geoespacial, Aguascalientes, Ags., Mexico
[2]Centro de Investigación en Matemáticas, Monterrey, Nuevo León, Mexico

ABSTRACT

Social media data is currently the main input to a wide variety of research works in many knowledge fields. This kind of data is generally multimodal, i.e., it contains different modalities of information such as text, images, video or audio, mainly. To deal with multimodal data to tackle a specific task could be very difficult. One of the main challenges is to find useful representations of the data, capable of capturing the subtle information that the users who generate that information provided, or even the way they use it. In this paper, we analysed the usage of two modalities of data, images, and text, both in a separate way and by combining them to address two classification problems: meme's classification and user profiling. For images, we use a textual semantic representation by using a pre-trained model of image captioning. Later, a text classifier based on optimal lexical representations was used to build a classification model. Interesting findings were found in the usage of these two modalities of data, and the pros and cons of using them to solve the two classification problems are also discussed.

KEYWORDS

Multimodal Data, Deep Learning, Natural Language Processing, Image captioning.

1. INTRODUCTION

Nowadays a large amount of data is generated by users on the Internet, particularly on social media platforms. This shared information usually represents feelings and opinions about social events, people or products; thus, we can also find humorous, sarcastic, offensive, motivational content, among others. The type of data shared is usually unstructured, which means, they do not have a well-defined order or organization and the relationship between their characteristics or variables is not clearly determined. Examples of them are text, images, video or audio. When different types data are involved, one can say we have multimodal data. In fact, our experience as human being is multimodal because we see objects, hear sounds, feel the texture, smell odors, and taste flavors [1]. In this way, the shared information cannot be stored in a traditional relational data structures, since this cannot be treated as usual in relation SQL architectures.

Multimodal data has become very interesting in the machine learning (ML) research community due to many classification problems could be solved by using more than a single modality of the data [2][3]. For instance, when the task is to classify sentiment on twitter users, we would expect that using images and text could be better than using only text. One of the main challenges of using multimodal data relies in the search for the optimal representation of all modalities of

David C. Wyld et al. (Eds): VLSIA, INWES, CSIP, BMLI - 2021
pp. 35-47, 2021. CS & IT - CSCP 2021 DOI: 10.5121/csit.2021.112105

information combined, where a supervised or non-supervised ML model can be used efficiently and effectively. Sometimes, one could think that the more information the better, but it is not always the truth. The quality and singularity of the data are more important usually.

When social media data is analysed, many insights and applications could be helpful, for instance, determining the attitude over products [4], politics [5], health care [6], events [7], or comments as protests [8]. In many of these problems, not only the text is shared, but also images or videos, and dealing with different modalities of information becomes more difficult but the idea of using the multimodal version of the data is very attractive most of the time.One of the most used platforms of social media adopted by the scientific community is Twitter, not only by computer science researchers, but also from other research areas such as economics, health, environmental studies, and many more.

With the aim of analysing if using two modalities of information for classification problems is better or not, in this paper we address two classification problems where images and text are available. We present a comparison of the solutions obtained by using a support vector machine (SVM) as the classifier and a specific representation of the input, that means of only text, only images, and a combination of both. This comparison provides a measure of how the performance increased or decreased according to the modality of information we used and also with the problem tackled.

Specifically, the contributions of the presented manuscript are the following:

- We obtained a semantic representation of images based on Deep Learning (DL) models, specifically, an Image Captioning model is implemented with visual attention to describe images data. That is, translate the image to text in a semantic way.
- The use of multimodal information in classification tasks is evaluated with experiments using only text, only images, and a combination of both.
- The evaluation was done with two difficult classification problems, tackled usually only with text data: user profiling and memes polarity classification.
- We made an implementation for the Spanish language based on an image captioning corpus available in English using automatic translation tools. In this way, we can use our proposed model for classification problems in Spanish and English languages.

The paper is organized as follows. Related Work Section describes the related work. In Datasets Section the data we used as well as the two classification problems we tackled are described. Methodology Section details all the steps of the proposed methodology, that means, image captioning problem, text representation, and text and image representation for the classification problems. Experiments and results are given in Experiments Section, and finally, the main findings and conclusions are described in Conclusions Section.

2. RELATED WORK

In recent years, there has been an increasing number of research efforts which aims to model and combine the information from each modality of data to tackle specific tasks. State of the art (SOTA) results for many tasks related to multimodal data are based on deep learning (DL) architectures and natural language processing (NLP) models. First research works on that field, combines information from images and text, and prior to the popularization of DL models, most of the research relied on feature engineering on both modalities of information, such as filters (e.g., Gabor or Sobel) for images, and lexical features based on n-grams for text, which were combined based on heuristic rules [9] or simple concatenation of the single-modality features [10]. In both researches, it was shown that the performance of the final classifiers improved

significantly when multimodal data was included instead of using just one modality. Popular tasks where image and text has been used are sentiment analysis and author profiling, promoted by the PAN@CLEF [11] [12] and SemEval [13] where SOTA results were achieved using deep encoder-decoder architectures, for both, images and texts, where specialized neural networks are used to extract features automatically, such as pre-trained convolutional neural networks, CNN [14] [15] for images and recurrent neural networks based on long short-term memory, LSTM [16], gated recurrent units, GRU [17] or transformer-based architectures such as BERT [18], for text sequences, where word embeddings are the common vector representation for words [19], [20], [13]. Once obtained the representation for each modality, they are combined generally with early or late fusion techniques [21], [22]. After that, a classifier such as neural network (NN), SVM or random forest (RF) is used, where the input is the whole multimodal representation.

There are other interesting applications where multimodal data has been used. In [23], a classifier for skin lesion is proposed based on multiple imaging modalities (macroscopic and dermatoscopic) and patient metadata such as age, sex, location of the disease, change of lesion, among others. Features for images were obtained with two pretrained CNNs based on ResNet-50 architecture [24], which are concatenated with patient metadata, followed by late fusion with a fully connected NN with two-hidden layers and a five-class softmax output layer in order to combine all features and obtain the corresponding lesion category. The authors show that the multimodal classifier outperforms one based on a single image. Multimodal data has been used extensively in different tasks related to music collections such as genre and mood classification or information retrieval. In this case, it is assumed that there are three main modalities of information [25] which can be associated to musical items: editorial (date of production, genre, composer, country of origin, etc.), cultural, (knowledge produced by the environment or culture, gathered from user profiles, web scraping or collaborative filtering) and acoustic (beat, tempo, rhythm, energy, and music structure). Multimodal information has been used for multi-label genre classification in [26], including cover art images, spectrograms from audio signal and customer reviews, where DL and NLP techniques are used for each modality of data. In [27], a proposal for tag-based music retrieval based on metric learning is presented, where the main idea is to create a shared embedding space based on acoustic and cultural embeddings obtained from Mel spectrograms and a user-song interaction matrix, respectively, in such a form that similarities between music items can be obtained.

3. DATASETS

Several datasets were used in this work. First, the Flickr30K [28] is used for image captioning task, i.e., textual descriptions of images.

For testing the multimodal classification approach, two more datasets related with the two classification problems were considered. These datasets correspond to Memotion Analysis SemEval2020 competition [13], and CLEF author profiling competition [12].

Flickr30k

For the image captioning model, we used the Flickr30k corpus, which consists of 158,915 captions from multiple sources that describe 31,783 images. These captions are in English language and the image's content is focused on people involved in everyday activities and events. These images are obtained from Flickr, a website that allows you to store, order, search, sell and share photographs or videos online, through the Internet.
Figure *1* shows some examples from the Flickr30K dataset.

Gray haired man in black suit and yellow tie working in a financial environment.
A graying man in a suit is perplexed at a business meeting.
A businessman in a yellow tie gives a frustrated look.
A man in a yellow tie is rubbing the back of his neck.
A man with a yellow tie looks concerned.

A butcher cutting an animal to sell.
A green-shirted man with a butcher's apron uses a knife to carve out the hanging carcass of a cow.
A man at work, butchering a cow.
A man in a green t-shirt and long tan apron hacks apart the carcass of a cow while another man hoses away the blood.
Two men work in a butcher shop; one cuts the meat from a butchered cow, while the other hoses the floor.

Figure 1. Images with captions from Flickr30k

In order to obtain captions in the Spanish language, a new version of Flickr30K was generated. This new version is just a translation of the captions to Spanish, which allowed to train a model to generate captions in Spanish.It is important to have a model for the Spanish Language because one of the two classifications task used in this work contemplates a task in Spanish. In Figure *2* we show some captions translated into the Spanish language.

image	...	caption_spanish
1000268201.jpg	...	Un niño con un vestido rosa sube las escalera...
1000268201.jpg	...	Una niña con un vestido rosa en una cabaña de...
1000268201.jpg	...	Una niña subiendo las escaleras hasta su casa...
1000268201.jpg	...	Una niña que sube a una casa de juegos de mad...
1000268201.jpg	...	Una niña entrando en un edificio de madera
1000344755.jpg	...	Alguien con una camisa azul y un sombrero est...
1000344755.jpg	...	Un hombre con una camisa azul está parado en ...
1000344755.jpg	...	Un hombre en una escalera limpia la ventana d...
1000344755.jpg	...	hombre de camisa azul y jeans en escalera lim...
1000344755.jpg	...	un hombre en una escalera limpia una ventana
1000366164.jpg	...	Dos hombres uno con una camisa gris uno con u...
1000366164.jpg	...	Dos chicos cocinando y bromeando con la cámara
1000366164.jpg	...	Dos hombres en una cocina cocinando comida en...
1000366164.jpg	...	Dos hombres están en la estufa preparando com...
1000366164.jpg	...	Dos hombres están cocinando una comida

Figure 2. Examples of translated text to Spanish from the Flickr30k dataset

Memotion Analysis

Memotion analysis [13] is an academic competition organized by SemEval (International Workshop on Semantic Evaluation) in 2020. The purpose of this competition was to provide a dataset including text and image, to tackle the problem of meme's classification, i.e., to detect for example when a meme is offensive or not. Meme's classification is far more difficult than the classification of text (for instance tweets), because a meme contains both, text and images to communicate its content. Nowadays there is not much attention to sentiment analysis in memes. When this task was launched, the objective was to attract the attention of the scientific community towards the automatic processing of memes shared on the internet on platforms such as Facebook, Instagram, and Twitter. Memes are difficult to deal with because they are often derived from our social and cultural experiences, such as television series or popular cartoon characters, and, as is stated in [13], "*these digital constructs are so deeply ingrained in our internet culture that to understand the opinion of a community, we need to understand the type of memes it shares*".

The dataset includes 7,000 images for training and 1,000 images for testing. For each meme, the text contained in the image is also shared. Then, we have the original image of the meme and the correct text extracted from it.
Figure *3* shows some examples of the Memotion Analysis dataset. In these examples, one has the image and the associated text.

Figure 3. Example of memes from the Memotion Analysis dataset

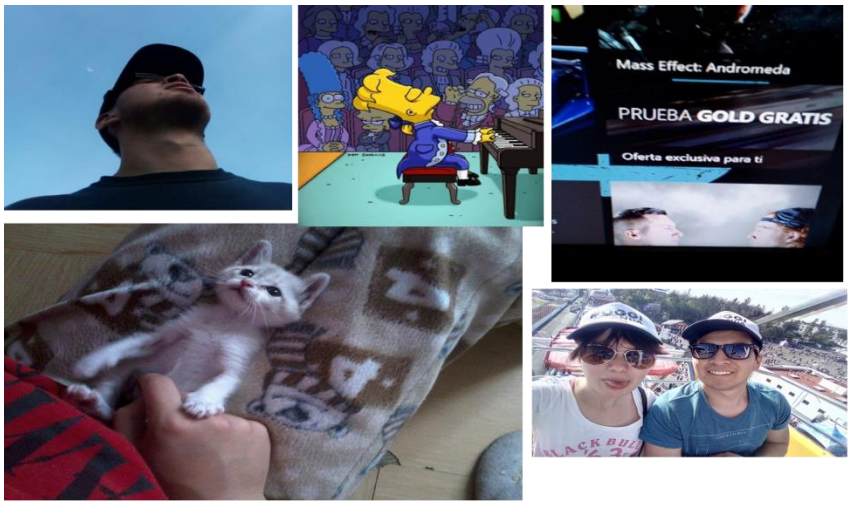
Author profiling

Author profiling is a task proposed by PAN[1], which is a series of scientific events and shared tasks on forensic digital text and stylometry and which in turn belongs to CLEF[2](Conference and Labs of the Evaluation Forum).Author profiling task tries to distinguish between classes of authors studying their sociolect aspect, that is, how language is shared by people. This helps to identify characteristics such as gender, age, native language, or personality type. For this task, the focus is on social networks, since it is of great interest to get insight of how everyday language reflects basic social and personality processes. Specifically, the data is generated by Twitter users.

The data used in this work is from the 2018 competition, which focused on gender identification on Twitter, where text and images are shared as sources of information. The languages addressed were English, Spanish, and Arabic. But, for this work, only the data in the Spanish and English languages were considered. In this dataset, there are 100 tweets and 10 images per user, as well as the gender for each of the 3,000 authors, in each language.

Figure *4* shows some examples of the data shared for this task [12].

[1] https://pan.webis.de/
[2] http://www.clef-initiative.eu/



[Bellos recuerdos xD https://t.co/wh07ILJwvJ]]></document>
[@Jokereando ojalá tuviera tu imaginación, yo solo veo una puta esponja :c]]></document>
[@Jokereando XD yo si le atiné cuando me preguntaron, bueno, dije esponja en forma de huevo, pero Masha? Cuerpos sin piel? Xdddd.]]></document>
[Era alguien con quien siempre podía hablar de lo que fuera, así sin miedo a ver su mensaje y no saber qué contestar, siempre había algo.]]></document>
[No hay nada más falso que un "jajaja" contestado tres días después, ya vio tu msg, ya lo ignoró, y ya se aburrió de que nadie más le habla.]]></document>
[El boxeo de sombra es lo más difícil del mundo, siempre que intento pegarle, la maldita se quita y le doy a la pared :'c #fitness]]></document>
[Mi vida es tan triste que hasta la de soporte técnico de Xbox me decepciona :'v #XboxOne #GoW4 #microsoft https://t.co/kOsl2yepMA]]></document>
[Justo después de comprar gold y canjear el código, hijo de tu puta madre #BillGates ¿cuántos más? https://t.co/Ylzf13N1V2]]></document>
[@aquilesNfierro @PeaPoblano @borrel0000 que sigue siendo mejor que cualquier cosa que el PRI pueda ofrecer...]]></document>
[¿Será que todos somos igual de ojetes y que la diferencia sea el nivel de hipocresía de cada quien?]]></document>
[Tal vez me tratarías con más respeto si supieras que vendí mis dólares para poder salir contigo, sabiendo que iban a subir así de cabrón. 🐱]></document>
["Estamos habituados a que los hombres hagan burla de lo que no entienden, y murmuren a la vista de lo bueno y lo bello". #Goethe]]></document>
["El mundo de los espíritus no está cerrado; tu sentido está obtuso, tu corazón está muerto." #Fausto #Goethe]]></document>
[Y lo peor es que de seguro muchos de estos vándalos pendejos saben que sólo afectan al pueblo, pero igual quieren sus cosas grátis.]]></document>
[No es bueno dejarse arrastrar por los sueños y olvidarse de vivir.]]></document>
[No es que te deseé el mal, pero ojalá te salgan puros Pinsir en tus huevos de 10 km.]]></document>
[@Jokereando cómo se llama esa serie?]]></document>
[@lopezobrador_ Es triste ver cómo la gente comenta tanta mierda, ni siquiera críticas constructivas, tienen el país que merecen y nos llevan]]></document>
[El que vence modificando su estrategia, según la condición de su enemigo, debe ser tomado como un experto en #ElArteDeLaGuerra.]]></document>
[Si no me quisiste cuando manejaba un Platina, no me busques ahora que tengo un Camaro B| ... Es lo que diría si tuviera un Camaro :'v xd]]></document>
[Ese Euler andaba en todo, me pregunto si tenía amiguitos.]]></document>
[@Terafobia si la frase es buena, ¿qué importa? Habrá quien la use porque YOLO y quien la use porque sí la entiende xd.]]></document>
[¿Por qué las mujeres son tan mierda?(Excepto Dalia mi BF) ¿Deberia probar la promiscuidad? ¿La homosexualidad? ¿O la promiscuidad bisexual?]]></document>
[¿Han tenido esa sensación de que están en un lugar y/o momento en el que no deberían estar?]]></document>
[No busques perfección, pero sí a alguien que mínimo te respete. Mereces algo mejor. ;)]]></document>
[Ojalá me gustara practicar el deporte tanto como me gusta verlo #Río2016]]></document>
[Tu infancia terminó cuando te diste cuenta que las palabras "heroína" y "erotico" no tienen que ver con superheroes, o que si...]]></document>
[No es lo mismo que te digan: "Voy a estar para lo que necesites", a: "Estaré siempre para ti, para lo que sea".]]></document>
[Dios bendiga al código abierto! :'D]]></document>
[Tan culpable es el que inicia el problema como el que no hace nada por resolverlo.]]></document>
[Antes de quejarte de la actitud de alguien más, mírate a ti.]]></document>
[No siempre se puede fingir una sonrisa para alegrar a alguien, a veces lo mejor que puedes hacer es no mostrarle lo agüitado que estás.]]></document>
[Perder duele, pero cada vez menos]]></document>
[A veces pienso que los sentimientos son nada más que un estorbo.]]></document>
[No importa que nadie lo note, tú sabes que hiciste algo bueno por alguien más :'(]]></document>
[Una simple vicisitud o una completa tribulación, la mayoría de las veces puedes decidir como ver a tus problemas.]]></document>
[No soy adepto a expresar mis emociones ni decir lo que pienso abiertamente... y no me harás empezar hoy. A menos claro, que esto cuente :P]]></document>
[Lección de vida, cortesía de House: "Trata a todo el mundo como si tuviera síndrome de Korsakoff". #TodosMienten]]></document>
[#CocaColaZeroAntártida tocan en la antartida... no tocan trapped under ice.... me dejaron con las ganas]]></document>
[#CocaColaZeroAntártida es impresionante, me encanta, pero me pregunto porque siempre tienen que tocar las mismas]]></document>
[@Jokereando no se viejo tengo tarea y mi gatita sufre de ataques de nuevo]]></document>
[@Jokereando estoy aquí los fines de semana, holis]]></document>
[@CartoonMX yo viendo la peli de la fusion de goku y vegeta a las 3am y a media pelicula ponen el final de los guerreros de plata!!!]]></document>
[@Jokereando XD pues ven por tus cosas]]></document>
[@Arturo_582 que es eso?? XDD]]></document>
[@arturo_582 se fue mi internet, yo imprimo , sabes el apellido de poulet?]]></document>

Figure 4. Examples of images and text shared in the author profiling 2018 dataset

4. PROPOSED METHODOLOGY

The proposed methodology consists of two steps. First, we obtain a textual semantic representation of images by training an image captioning model with the Flickr30K dataset on English and Spanish languages. In the second step, we fuse vector representation (obtained with NLP techniques) of the two modalities of our data by concatenating previously the text information from our data and the one obtained in the first step. On this shared representation, we train two classification models based on SVM to solve the memes classification problem and the user profiling task.

Figure *5* shows a schematic overview of the methodology, and the details are described in the following subsections.

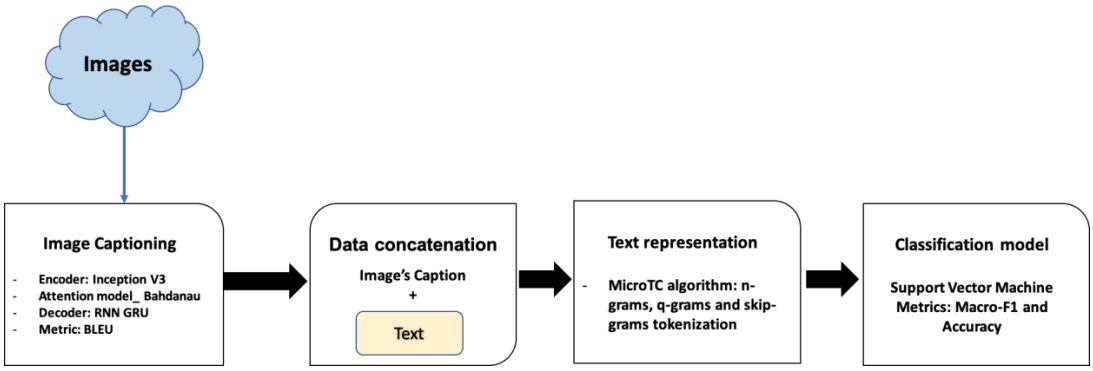


Figure 5. Scheme of our Proposed methodology

Image captioning step

Image captioning is the task which attempts to describe, in a semantic way, an image content. SOTA results for this task are achieved by deep encoder-decoder architectures, as the one we used, which is based on [29]. The encoder consists on a pre-trained CNN based on the Inception-v3 architecture [30] with visual attention [31], in order to relate, or "align" some specific objects of an image with its corresponding text descriptions given in the training data. The output of the encoder is a vector representation of the image, which in turn, is the input of the decoder, which learn to generate an output sequence which is the textual description of the image. In our case, the decoder is based on a recurrent neural network (RNN) with gated recurrent units (GRU). The image captioning architecture is shown in

Figure 6. As we said before, our training corpus for this task is Flickr30k dataset, described in Flickr30k Section.

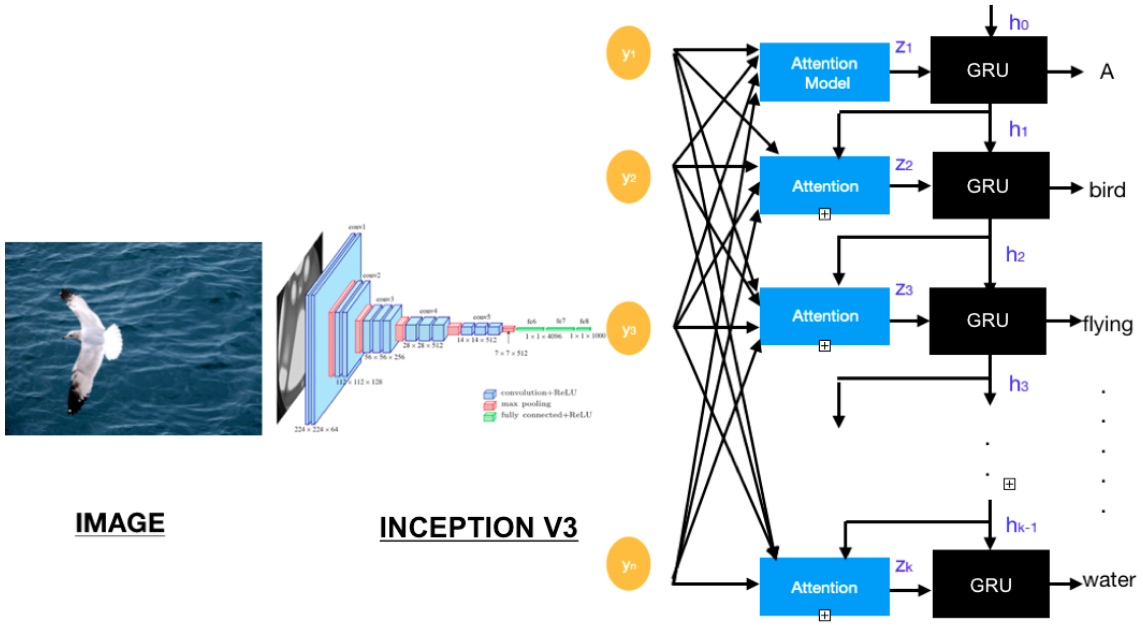


Figure 6. Image Captioning model[3].

[3] Source: https://medium.com/swlh/image-captioning-using-attention-mechanism-f3d7fc96eb0e

Multimodal fusion and classification

Once we have the text which corresponds to the image description and the original text provided in the dataset, multimodal fusion is carried out by learning a vector representation of the concatenated text from both modalities of data. To this end, we used μTC algorithm [32], which is defined as a minimalist text classifier based on SVM, robust to any language and domain. The main idea in [32] μTC is searching for an optimal text representation based on a set of text transformations such as noise deleting, normalization, and tokenization, among others. The optimal representation is the one which has a good performance in a given classification problem. The procedure can be viewed as a combinatorial optimization problem wherein each algorithm iteration, the performance of the parameter's configuration is measured trying to select a better configuration in the next step until the best possible solution is reached. As final step, the vector generated with μTC is used as the input to a SVM classifier. As performance measures, we used macro-F1 and Accuracy metrics.

5. EXPERIMENTS AND RESULTS

In this section, several experiments are presented. First, the performance of the image captioning task is assessed. In this case, we used the bilingual evaluation understudy (BLEU) metric (please see [33] for technical details), which measures the grammatical composition of sentences with n-grams in order to evaluate if the candidate caption obtained, captures the meaning of the reference caption given in the dataset. A key aspect in BLEU is the n-grams considered, which refer to a sequence of words within a window, where n represents the size of the window. For example, for the sentence "yesterday I went to the park to run"the unigram (n=1) represent each word, while for n = 2 we have bigram: "yesterday I", "I went", "went to", "to the", "the park", "park to", "to run". Thus, in BLEU the n-gram of the candidate caption is compared with the n-gram of the reference caption. It is worthwhile to mention that BLEU does not consider the position of the n-grams in the text but the number of matches. In the case of our image captioning model trained with Flickr dataset, the BLEU score obtained based on a 20% testing dataset and unigram were 42 and 38.7 for Spanish and English languages, respectively, which are considered as high quality and good results. In

Figure 7 *and*

Figure *8* we show some representative examples of the image captioning model results. We can see that good results are obtained in both languages, but outstanding results are observed for Spanish.

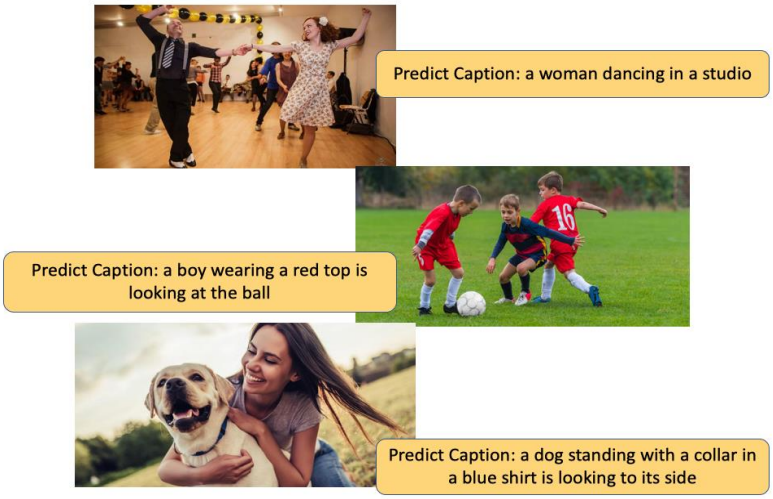


Figure 7. Example of captions in English generated by the image captioning generated model

Figure 8. Example of captions in Spanish generated by the image captioning with our generated dataset (Spanish dataset)

Once we obtained our model for image captioning, we proceed to assess the meme classification and author profiling tasks. For the evaluation, we used macro-F1 and Accuracy metrics.In the evaluation of meme's classification, we tackled two out of three tasks related to the Memotion analysis contest. In task A (polarity classification), the objective is to classify a meme content as positive, negative, or neutral, meanwhile in Task B (humour classification), the objective is to classify a meme as sarcastic, humorous, offensive or motivator, and further, a meme can be classified in more than one category, making this task very difficult.

Two baselines were provided for the Memotion analysis competition [13]. For task A, macro-F1 was 0.2176, and for task B 0.5118. Furthermore, the best results obtained by competitors of the Memotion analysis were 0.3546 for task A, and 0.5183 for task B, both in macro-F1 (see [13] for all results).

Table 1shows our results for meme classification in both tasks. In Task A, we can see very good results in the training stage, but not so good in testing. By using only text information, a macro-F1 of 0.955 is obtained but dropped dramatically to 0.854 in testing. This could be for overfitting in the training stage. When we used Text + Caption, the performance was similar, reaching a 0.976 of accuracy, higher than using only text, but in testing the performance also dropped. In the case of Task B, lower results were reached using Text or Text + Caption. Although we obtained better results in Task-A versus the best result of the competition, in Task-B we achieved a very

lower result compared again with the best of the competitors. Our objective is to analyze if our generated caption improves the classification performance, in this case, we can see this has not happened.

Table 1. Memes classification results (macro-F1)

Task	Data	Training	Test	baseline	Best score competition
Task-A	Text	0.955	0.854	0.2176	0.3546
Task-A	Text + Caption	0.976	0.740		
Task-B	Text	0.437	0.380	0.5118	0.5183
Task-B	Text + Caption	0.444	0.369		

For the author profiling task, we used the dataset reported in [12] for English and Spanish languages. Because this competition was held in 2018, we had no access to the baseline nor testing dataset, so, we split the dataset into 80% for training and 20% for testing. For reference, we show in Table *2* the best results for Spanish and English achieved in the competition as was published at that time. The metric used here is accuracy.

Table 2. User profiling contest best results (accuracy)

Language	Data	Training	Test
Spanish	Text	-	0.8200
Spanish	Text + Caption	-	0.8200
English	Text	-	0.8221
English	Text + Caption	-	0.8584

The results reached with our proposed methodology are shown in Table 3. Here, it can be seen the perfect results in training for both languages, but a lower accuracy for test. In both cases, a better result was obtained by considering multimodal data, Text + Caption. Nevertheless, we consider that there is not enough evidence to demonstrate the advantage of including image information by means of its corresponding caption.

Table 3. User profiling results (Accuracy)

Language	Data	Training	Test
Spanish	Text	1.00	0.823
Spanish	Text + Caption	1.00	0.833
English	Text	1.00	0.735
English	Text + Caption	1.00	0.738

As an analysis, we can state that for the meme classification task, a better performance was only obtained in Task A, compared to the baseline reported in the competition. However, it was observed that the classification was better when using only the textual information in both tasks (A and B), and by combining both modalities of data (Text + Caption) it was not possible to achieve good generalization. Then, we could conclude that the descriptions of the images did not improve significantly the results regarding this task. One of the possible reasons for this situation is the composition of the dataset for this task since the sentiment contained in the meme is mainly identified with the textual message of the image, in addition, lower performance may occur in the model by adding the information of the images because the same image is used to create memes with different sentiments, i.e., the same image with different text could be an opposite sentiment perspective.

Respecting to the author profiling task, we obtained better performance when using the textual information combined with the image's description, compared to the performance obtained by classifying with only the user's textual information, in both languages. When we analyse the results reported in this competition, we found that the best result was achieved by using only text than using only images.In our case, although the improvement when using both modalities of information seems to be not so significant, an issue that must be taken into account is that for each user there were around 100 tweets and only 10 shared images and, based on the length of the tweets versus the descriptions generated for each image, the textual information of the added images represents only 9% of the final content for each user, resulting in a significant unbalance in the data modalities.

In both tasks, there are some issues regarding their datasets such as data distribution, topics, categories definition, almost the same data could represent different categories or classes, among others.Maybe, the extension of this analysis with more data could show a better understanding on the relevance of image captioning in this kind of problems.

6. CONCLUSIONS

Our main objective was to demonstrate if using the image captioning approach to use multimodal data could improve the classification results in both tasks, Meme classification, and User profiling.In general, the methodology applied for the classification of multimodal data had a good evaluation in both tasks but not significantly outstanding result to state that using captions in multimodal data improves the performance of the two classification problems tested.

Even so, to conclude on its performance compared to other approaches, we think it is necessary to apply in tasks with balanced data types, i.e., with the same number of images and texts for each sample, in order to analyse if the semantic description of the images provides significant information in the representation of the data. Our methodology did not obtain a good generalization with the description of the images in the classification for new data in the tasks addressed, this is reflected by excellent results in training that decrease in the test dataset, which is a topic of interest as future work, also, it is in our interest to explore another DL architectures, such as those based on transformers.

As future work, more fusion techniques could be applied to both, images and text data. Also, it will be interesting to test our methodology with more adequate datasets to state more solid conclusions and produce more accurate image captioning models to improve the image semantic descriptions.

ACKNOWLEDGEMENTS

This research is supported by the project A1-S-34811 of Basic Science grant by the National Council of Science and Technology (CONACyT) from Mexico.

REFERENCES

[1] Baltrušaitis, T., Ahuja, C., &Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. IEEE transactions on pattern analysis and machine intelligence, 41(2), 423-443.

[2] Kumar, A., & Garg, G. (2019). Sentiment analysis of multimodal twitter data. Multimedia Tools and Applications, 78(17), 24103-24119.

[3] Alam, F., Ofli, F., & Imran, M. (2018, June). CrisisMMD: Multimodal twitter datasets from natural disasters. In Proceedings of the International AAAI Conference on Web and Social Media (Vol. 12, No. 1).

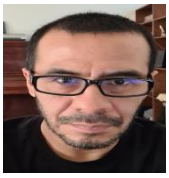
[4] Das, T. K., Acharjya, D. P., & Patra, M. R. (2014, January). Opinion mining about a product by analyzing public tweets in Twitter. In 2014 International Conference on Computer Communication and Informatics (pp. 1-4). IEEE.
[5] Jungherr, A. (2014). Twitter in politics: a comprehensive literature review. Available at SSRN 2865150.
[6] Sinnenberg, L., Buttenheim, A. M., Padrez, K., Mancheno, C., Ungar, L., & Merchant, R. M. (2017). Twitter as a tool for health research: a systematic review. American journal of public health, 107(1), e1-e8.
[7] Nichols, J., Mahmud, J., & Drews, C. (2012, February). Summarizing sporting events using twitter. In Proceedings of the 2012 ACM international conference on Intelligent User Interfaces (pp. 189-198).
[8] Segerberg, A., & Bennett, W. L. (2011). Social media and the organization of collective action: Using Twitter to explore the ecologies of two climate change protests. The Communication Review, 14(3), 197-215.
[9] Kalva, P., Enembreck, F., &Koerich, A. (2007, September). Web image classification based on the fusion of image and text classifiers. In Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) (Vol. 1, pp. 561-568). IEEE.
[10] Kannan, A., Talukdar, P. P., Rasiwasia, N., &Ke, Q. (2011, December). Improving product classification using images. In 2011 IEEE 11th International Conference on Data Mining (pp. 310-319). IEEE.
[11] Rangel, F., Rosso, P., Potthast, M., Trenkmann, M., Stein, B., Verhoeven, B., &Daelemans, W. (2014). Overview of the 2nd author profiling task at pan 2014. In CEUR Workshop Proceedings (Vol. 1180, pp. 898-927). CEUR Workshop Proceedings.
[12] Pardo, F. M. R., Montes-y-Gómez, M., Potthast, M., Stein, B., Ferro, N., Nie, J. Y., &Soulier, L. (2018). Overview of the 6th Author Profiling Task at PAN 2018: Cross-domain Authorship Attribution and Style Change Detection. In CLEF 2018 Eval Labs Workshop–Work Notes Pap (Vol. 10, p. 14).
[13] Sharma, C., Bhageria, D., Scott, W., PYKL, S., Das, A., Chakraborty, T., ... &Gamback, B. (2020). SemEval-2020 Task 8: Memotion Analysis--The Visuo-Lingual Metaphor!.arXiv preprint arXiv:2008.03781.
[14] LeCun, Y. (1989). Generalization and network design strategies. Connectionism in perspective, 19, 143-155.
[15] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.
[16] Hochreiter, S., &Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.
[17] Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., &Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.
[18] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.
[19] Takahashi, T., Tahara, T., Nagatani, K., Miura, Y., Taniguchi, T., & Ohkuma, T. (2018). Text and image synergy with feature cross technique for gender identification. Working Notes Papers of the CLEF.
[20] Álvarez Carmona, M. Á., VillatoroTello, E., Montes y Gómez, M., &Vilaseñor Pineda, L. (2020). Author profiling in social media with multimodal information. Computación y Sistemas, 24(3), 1289-1304.
[21] Liu, K., Li, Y., Xu, N., & Natarajan, P. (2018). Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730.
[22] Ramachandram, D., & Taylor, G. W. (2017). Deep multimodal learning: A survey on recent advances and trends. IEEE signal processing magazine, 34(6), 96-108.
[23] Yap, J., Yolland, W., &Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. Experimental dermatology, 27(11), 1261-1267.
[24] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).
[25] Pachet, F. (2005). Knowledge management and musical metadata. Idea Group, 12.

[26] Oramas, S., Barbieri, F., Nieto Caballero, O., & Serra, X. (2018). Multimodal deep learning for music genre classification. Transactions of the International Society for Music Information Retrieval. 2018; 1 (1): 4-21.
[27] Won, M., Oramas, S., Nieto, O., Gouyon, F., & Serra, X. (2021, June). Multimodal metric learning for tag-based music retrieval. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 591-595). IEEE.
[28] Young, P., Lai, A., Hodosh, M., & Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. Transactions of the Association for Computational Linguistics, 2, 67-78.
[29] Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., ... &Bengio, Y. (2015, June). Show, attend and tell: Neural image caption generation with visual attention. In International conference on machine learning (pp. 2048-2057). PMLR.
[30] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., &Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2818-2826).
[31] Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.
[32] Tellez, E. S., Moctezuma, D., Miranda-Jiménez, S., & Graff, M. (2018). An automated text categorization framework based on hyperparameter optimization. Knowledge-Based Systems, 149, 110-123.
[33] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics (pp. 311-318).

AUTHORS

DANIELA MOCTEZUMA is a researcher at the Research Center on Geospatial Information Sciences (CentroGEO) since 2014. She received her Ph.D. in Computer Sciences from Rey Juan Carlos University, Madrid, Spain in 2013. Her research interests include machine learning, computer vision, natural language processing, intelligent video surveillance systems, and remote sensing.

VÍCTOR MUÑIZ received a PhD in computer science at Research Center in Mathematics (CIMAT) in Guanajuato, Mexico. He is working at CIMAT Monterrey, Mexico, in liaison and research projects. His research interests are Machine Learning, Natural Language Processing and spatio-temporal models.



JORGE GARCIA Jorge Sánchez García received a bachelor degree in mathematics from Universidad Autónoma de Nuevo León (UANL) and a master degree in statistical computing from Research Center in Mathematics (CIMAT). His research interest is computer vision.