DANKMEMES for EVALITA 2020
Task Guidelines

Giulia Giorgi, Ilir Rama, Martina Miliani,
Guido Anselmi, Gianluca E. Lebani

29th May 2020

Contents

1 Task(s) description

DANKMEMES (multimoDal Artefacts recogNition Knowledge for MEMES) is the first EVALITA task for meme recognition and hate speech/event identification in memes. The DANKMEMES shared task is articulated into three subtasks and it is open to everyone from industry to academia. Participants can choose to take part to all or to some of the subtasks, which are:

1.1 Subtask 1 - Meme Detection

The first task consists in a binary classification, where systems have to predict whether an image is a meme (1) or not (0).

1.2 Subtask 2 - Hate speech Identification

The second task consists in a binary classification, where systems have to predict whether a meme is offensive (1) or not (0). Following the definition by [Zam+19], an offensive meme contains any form of profanity or a targeted offense, veiled or direct, such as insults, threats, profane language or swear words.

1.3 Subtask 3 - Event Clustering

The goal of the third subtask is to cluster a set of memes that may be or may be not related to the 2019 Italian government crisis into the following five event categories:

0. residual category reserved for memes that do not fit in any of the other classes;
1. the beginning of the government crisis;
2. the beginning of consultations involving political parties and Conte's Senate speech;
3. Giuseppe Conte is called by the President Mattarella to form a new government;
4. the 5SM holds a vote on his direct democracy platform, Rousseau.

Participants' goal is to apply supervised or unsupervised techniques to cluster the memes, so that memes pinpointing to the same events are put in the same cluster. Given that, participants are required to specify the algorithm type adopted to approach the task. In a first mandatory run, all participants must associate each item of the dataset to one of the events provided in the training set ("labelled run"). Participants approaching the task with clustering can submit an additional run of the system in which they can group items in an arbitrary number of clusters ("unlabelled run"). This means that for this task, there will be two different rankings.

2 Dataset

The DANKMEMES dataset is comprised of 2,361 images (for each subtask will be provided a specific dataset), automatically extracted from Instagram through a Python script aimed at the hashtag related to the Italian government crisis ("#crisidigoverno"). Posts have been collected along with their related metadata and then anonymized.

For each image, the following features have been collected:

- **File**: the name of the .jpg image file.
- **Date**: when the image has first been posted on Instagram.
- **Macro status**: refers to meme layouts and their relation to diffused, conventionalised formats called macros. The category has 0 and 1 as labels, where the value 1 represents well-known memetic frames, characters and layouts (e.g. Pepe the Frog).
- **Picture manipulation**: entails the degree of visual modification of the images. Non-manipulated or low impact changes are labeled 0 (e.g. the addition of a text or a logo). Heavily manipulated, impactful changes (e.g. images edited to include political actors) are labeled 1.
- **Visual actors**: the political actors (i.e. politicians, parties' logos) portrayed visually, regardless whether edited into the picture or portrayed in the original image.
- **Engagement**: the number of comments and likes of the image.
- **Text**: the textual content of the image has been extracted through optical character recognition (OCR) using Google's Tesseract-OCR Engine, and further manually corrected.
- **Meme**: binary feature, where 0 represents non meme images and 1 meme images. This is the target label for the first subtask.
- **Hate Speech**: binary feature only for memes. It differentiates memes with offensive language (1) from non offensive memes (0). This is the target label for the second subtask.
- **Event**: feature only for meme images, categorizing them according to 4 events (described in 1.3), plus a residual category labeled as 0. This is the target label for the third subtask.

The dataset also includes image embeddings. The vector representations are computed employing ResNet [He+16], a state-of-the-art model for image recognition based on Deep Residual Learning. Providing such image representations allows the participants to approach these multimodal tasks focusing primarily on its NLP aspects [KB14].

Participants are allowed to use external resources, lexicons or independently annotated data. Given that, although we provide ResNet image embeddings, participants can use any other image representations.

2.1 Development and Test Data

Along with the images and the image embeddings, participants will receive different training and test datasets for each subtasks. All datasets will be provided in utf-8 encoded comma separated ".csv" files, structured as follows:

Dataset for subtask 1 - Meme Recognition. The whole dataset counts 2,000 images, half memes and half not. We split the dataset into training and test sets, in a proportion of 80-20% of items. Table 1 represents the format of the training dataset. The test dataset will be provided without gold labels (i.e. without the "Meme" attribute) for testing purposes.

File	Engagement	Date	Manip.	Visual	Text	Meme
1.jpg	21,053	22/08/19	1	Conte	aiuto	0
56.jpg	114	22/08/19	0	Salvini	alle solite	1

Table 1: An excerpt from the dataset for the subtask 1.

Dataset for subtask 2 - Hate speech Identification. The whole dataset counts 1,000 memes. We split the dataset into training and test sets, in a proportion of 80-20% of items. Table 2 represents the format of the training dataset. The test dataset will be provided without gold labels (i.e. without the "Hate Speech" attribute) for testing purposes.

File	Engagement	Manip.	Visual	Text	Hate Speech
62.jpg	21,053	1	Conte	aiuto	0
10.jpg	12,572	0	Conte, Salvini	aspetta manca ancora la parte in cui parlo del lavoro di tua sorella	1
114.jpg	12,572	1	Salvini	merdman	1

Table 2: An excerpt from the dataset for the subtask 2.

Dataset for subtask 3 - Event Clustering. The whole dataset counts 1,000 memes. We split the dataset into training and test sets, in a proportion of 80-20% of items. Table 3 represent the format of the training dataset. The test dataset will be provided without gold labels (i.e. without the "Event" attribute) for testing purposes.

File	Eng.	Date	Macro	Manip.	Visual	Text	Event
43.jpg	21,053	22/08/19	1	1	Conte	aiuto	1
23.jpg	114	22/08/19	1	0	Salvini	alle solite	0
114.jpg	12,572	25/08/19	0	1	Salvini	merd-man	2

Table 3: An excerpt from the dataset for the subtask 3.

Image Embeddings. A .csv file containing the image embeddings will be provided for each training and test set of the three subtasks. In each file, vectors of dimension 2048 are associated with the corresponding image file name, whereas their elements are space-separated (see Table 4).

File	Embedding
1.jpg	0.10538975894451141 1.441086769104004 ... 0.10747165232896805
23.jpg	0.41991370916366577 0.49551108479499817 ... 0.0672694519162178
1453.jpg	1.7892037630081177 1.1843881607055664 ... 0.09812714904546738

Table 4: An example of an image embeddings file.

2.2 Distribution and Data format

The development data will be available by the **29th of May 2020**, whereas the test sets will be available by the **4th of September 2020**. Please note that the data will be made available in the Dataset section of our website: `https://dankmemes2020.fileli.unipi.it/`.

The download of data will be protected by a password. Contact us via email (dankmemesevalita@gmail.com) to get access to the files. Participants will receive the password for downloading the following material:

- three .csv files (one for each subtask) containing the metadata described in section 2;
- three folders (one for each subtask) containing the images in .jpg format;
- three .csv files (one for each subtask) containing the relative image embeddings 2.

All material is released for non-commercial research purposes only under a Creative Common license (BY-NC-ND 4.0). Any use for statistical, propagandistic or advertising purposes of any kind is prohibited. It is not possible to modify, alter or enrich the data provided for the purposes of redistribution.

3 Evaluation

For all tasks, the models will be evaluated with *Precision*, *Recall* and F_1 scores.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives}$$

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives}$$

$$F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

The performance of the multiclass classifiers proposed for the supervised setting of the third subtask will be computing the performance for each class and then averaging over all classes.

An additional ranking will be performed for clustering systems participating in the unsupervised setting of the third subtask, based on the Silhouette Coefficient [Rou87]. For each observation i, the Silhouette Coefficient is calculated as follows:

$$S(i) = \frac{b_i - a_i}{max(a_i, b_i)}$$

where a_i is the mean distance between i and all the other members of the same cluster and b_i is the mean distance between i and the items in the next nearest cluster. The measure used to evaluate the single models, the Silhouette over all samples, is computed as the average of each observation's Silhouette coefficient.

Different baselines will be used in the different subtasks:

- **Subtask 1 - Meme Detection:** the baseline is given by the performance of a random classifier, which labels 50% of images as meme.
- **Subtask 2 - Hate Speech Identification:** the baseline is given by the performance of a classifier labeling a meme as offensive when the meme text contains at least a swear word.
- **Subtask 3 - Event Clustering (labelled run):** the baseline is given by the performance of a classifier labeling every meme as belonging to the most numerous class (i.e. the residual one).
- **Subtask 3 - Event Clustering (unlabelled run):** the baseline is given by the performance of a model organizing memes solely on the basis of their publication date.

4 How to submit your runs

Submissions should be sent via email to dankmemesevalita@gmail.com using the subject "dankmemes [team-name]". Participants can pick any name of their choice and each team can submit **maximum two results files**.

Results should be submitted in CSV files named according to the following formalism: `dankmemes-taskname-teamname-run[1|2].csv`, where the allowed task names are "task1", "task2", "task3_labelled" and "task3_unlabelled". These files should contain the following data fields:

- "File": filename of the labelled image
- "Label": predicted label for each image

Participants willing to compete in the unlabelled version of the third subtask should submit an additional CSV file encoding the pairwise distances between the images of the test set. This file should be named as:
`dankmemes-task3_unlabelled_distances-teamname-run[1|2].csv`
and should contain the following data fields:

- "Image 1": filename of the first image
- "Image 2": filename of the second image
- "Distance": pairwise distance between the two images

References

[Rou87] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: *Journal of computational and applied mathematics* 20 (1987), pp. 53–65.

[KB14] Douwe Kiela and Léon Bottou. "Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics". In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2014, pp. 36–45.

[He+16] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.

[Zam+19] Marcos Zampieri et al. "Predicting the Type and Target of Offensive Posts in Social Media". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. 2019, pp. 1415–1420.