Sentiment Polarization in Online Social Networks: The Flow of Hate Speech

* Katerina Katsarou, Sukanya Sunder, Vinicius Woloszyn
Technische Universität Berlin
Berlin, Germany
{a.katsarou, sukanya.sunder, woloszyn}@tu-berlin.de

† Konstantinos Semertzidis
IBM Research Europe
Dublin, Ireland
konstantinos.semertzidis1@ibm.com

***Abstract*—The influence of sentiment polarization and exchange in online social networks has been growing and studied by many researchers and organizations worldwide. For example, the sentiments expressed in a text concerning a topic in the discussion tend to influence a community when a Twitter user retweets the original text, causing a chain of reactions within a network. This paper investigates sentiment polarization in Twitter, focusing on tweets with the hashtags #Coronavirus, #ClimateChange #Immigrants, and #MeToo. Specifically, we collect the tweets mentioned above and classify them into five categories: hate speech, offensive, sexism, positive, and neutral. In this context, we address the problem as a multiclass classification problem by using the pre-trained language models ULMFiT and AWD-LSTM, which achieved a F_{micro} of 0.85. Finally, we use the classified dataset to conduct a case study in which we capture the sentiment orientation during the network evolution.**

***Index Terms*—Hate Speech, Online Social Networks, NLP, Sentiment Analysis, Deep Learning, ULMFiT, Twitter.**

I. Introduction

The digital world is rapidly working towards AI-powered sentiment analysis that can help business owners to learn what people feel about their product, global movement campaigners how the audience is reacting, notify national security organizations of suspicious behavior, and in most recent times, assess possibilities for promising voters as a better political strategy. As an ongoing field of research in text mining, many studies are available relating to sentiment analysis on social media platforms and networks. The popularity of social media platforms has generated massive user content over 15 years, allowing the comprehensive and large-scale analysis of user behavior. In these years, studies like Bolen et al., 2009 [1] and Belkin et al., 2006 [2] have concluded that human emotions and sentiments do interchange within discussions mediated on online platforms such as forums, microblogging sites, and other channels. However, the true meaning of messages can be misunderstood at the micro-level due to the absence of non-verbal parts.

The role of sentiments and influencers in triggering discussions among users in online social media is investigated by Hillmann and Trier, 2012 [3]. Their work adopts a perspective that models users as nodes and exchange of messages as links within complex social networks.

The work in [4] analyses the sentiments expressed during the 2011 Egyptian Revolution. The Twitter platform was the data source for this study due to its position as an ideal destination for public opinion and soon became the forefront during the political uprising. The work mentions how the restriction of 140 characters meant that the tweets directly represented the emotions concerning the event. It was only 11 years after launch, in November 2017, that Twitter extended this restriction to 280 characters[1] to accommodate non-CJK languages. Despite this extension, it is widely believed that user messages related to any event closely represent their emotions.

Given the volume of social media posts generated every hour, manual classification of messages for hateful content is neither robust nor scalable; hence it has motivated researchers to find automated solutions. Following the work done in [5], deep neural networks effectively solve several language processing tasks such as part-of-speech tagging, sentiment analysis, and named entity recognition. The majority of publicly available datasets that are commonly used in hate speech detection classify the text as either offensive or not offensive, and some records exist with *racism* and *sexism* classifications. However, defining the classification of a tweet and formulating context-dependent foul language examples makes this task quite challenging. This is due to the inherent complexity of the natural language constructs such as different forms of hatred, different groups of targets, and synonyms that represent similar meanings.

In this paper, we address the challenges faced by traditional machine learning methodologies in sentiment classification. The challenges for Natural Language Processing (NLP) tasks escalate with the growth of hate speech against specific groups like immigrants, women, and climate activists. To this end, we present a transfer learning approach that uses semi-supervised learning on a pre-trained model for multiclass classification. We evaluate our approach by classifying tweets from Twitter that are related to topics such as immigrants, climate change, MeToo movement, and Coronavirus Pandemic into the following five different categories: i) hate speech, ii) offensive, iii) sexism, iv) neutral and v) positive. Following text classification, we focus on capturing the sentiment orientation during the network evolution by seeing the data as graph networks that contain communities in which hate speech flows between

[1] https://en.wikipedia.org/wiki/Twitter

different communities.

In summary, in this paper, we make the following contributions:

- Evaluation of ULMFiT using six different public labeled datasets, which outperformed the related work. We manage to have a fined-grained approach that performs five classes classification for addressing the problem of hate speech and achieves highly accurate predictions;
- The framework also is tested on five other datasets from the related work, and there are comparisons on the evaluation results. The harvested evaluation criteria are outperforming those obtained by the pioneering classifiers of the literature;
- The classified tweets belong to four different networks: i) #Immigrants, ii) #MeToo, iii) #Coronavirus, and iv) #ClimateChange and for each of them, we analyze the sentiment orientation during the evolution of the networks. Louvain, Weakly connected components, label propagation and asynchronous label propagation algorithms are implemented in order to detect communities within these networks. These networks' evolution and their characteristics are monitoring over time.

The rest of this paper is structured as follows. In Section II, we present the related works and their limitations. In Section III, we describe the Proposed Framework, and in Section IV, we present the results of our extensive experimental evaluation. Section V concludes the paper.

II. Related Work

Three main research areas are associated with our work, namely, hate speech detection, sentiment analysis, and graph theory. In [6] the proposed framework identifies hate speech against women. The authors performed three experiments with data extract from Twitter, being two addressed to detection of misogyny and one for sexism. Their approach employed lexical and stylistic features such as a lexicon containing vulgarities and another lexicon with words related to sexuality, where the relevance of the words is calculated with Information Gain. A Support Vector Machine (SVM) classifier was used with accuracy of *0.76*, *0.8* and *0.89* for the three corpora. Another example is the approach from Sharma et al., [7] that created an ontological classification of harmful speech based on the degree of hateful intent. In their experiments, the authors compared the performance of Naive Bayes, SVM, and Random Forest Classifiers. The SVM achieved the best accuracy of *0.76*.

Recently, sentiment analysis with contextual semantics has been used in Twitter to segregate discussions by topics and subtopics. Apart from Twitter, other social platforms infamous for fuelling discussions around controversial topics have been used for sentiment analysis. For instance, Gab in [8], Wiki comments in [9], and Reddit in [10]. In [8], 21M posts of 341K users over 20 months were collected. Then, DeGroot's information diffusion model was used to present the iterative nature of hate speech propagation among users. The findings show that the users are densely connected and produce *18.7%* of the overall posts in Gab. Wulczyn et al. [9] use 100k human-labeled comments and 63M machine-labeled ones and Logistic Regression (LR) and Multi-layer Perceptrons (MLP) for the binary classification of the comments into attacking and non-attacking. They evaluate the results with Area Under Curve (AUC) and Spearman rank correlation. The authors acquired an AUC of *96.6%* and *68.2%* for Spearman with MLP classifier. Tsantarliotis et al. [10] analyse 555,332 comments from Reddit and create metrics such as *TVDiff*, *TVRatio* and *TVRank* for predicting if a post is vulnerable to trolls or not. A classifier is used with features such as content, authors, and history, and the results are evaluated with metrics such as accuracy, precision and recall, and AUC. The best AUC for *TVDiff* is 0.79, for *TVRatio* 0.82 and for *TVRank* 0.83.

Several graph theory researchers like Fornacciari et al., [11] apply follower relationships, deriving sentiments of the user based on the sentiments of their tweets. The case study in this work is the social network of the #SamSmith channel (the singer who won four awards at the Grammy Awards 2015). The authors classify the tweets into objective and subjective posts. Only subjective posts contain sentiment polarity (positive or negative). Sixty channels were explored in Twitter, and the polarity classifier gave *76.5%* accuracy and the subjectivity classifier *79.5%*. Another method fuelling tweet engagement is the tweet-retweet relationship, used in [12] to predict the *retweetability* of a tweet, using content-, social-, author- and tweet-based features. Then, the authors used classification frameworks such as SVM, J48, and logistic regression, with J48 performing better (with recall, precision, and F1 score equal to 0.84, 0.825, and 0.832). In [13] target-dependent features are also used to leverage the relationships among multiple tweets to enhance the sentiment classification of the text. Speriosu et al., [14] compare a lexicon-based baseline, a classifier based on max entropy and label propagation. These studies have, to some extent, used additional tweet information. Founta et al. [15] propose a deep-learning-based approach, where they use two classification paths, one for the text-based features and one with the non-sequential metadata features. As metadata features, the authors use content-based, user-based, and network-based features. Then, they use different techniques for training their classifiers, such as transfer learning with having both classifiers pre-trained separately, transfer learning with fine-tuning, and combined learning with interleaving. All the approaches above give an AUC between *92%* and *98%* for the used datasets. In the work of Cao et al. [16], a deep-learning-based architecture was proposed that consists of CNNs, LSTMs, and attention layers and utilizes word embeddings, sentiment, and topical information, resulting in an F1 score that varies from 0.72 to 0.89 depending on the dataset. Rizoiu et al. [17], present different transfer learning-based variations that use ELMO embeddings and Bidirectional LSTMs for two different tasks. The first task aims to distinguish harmless messages from racist or sexist and the second one from hateful or offensive with F1 scores of 0.78 and 0.72, respectively. Finally, Perifanos et al. [18] propose a model based on BERT and Residual Neural Networks with

transfer learning. The authors used hateful, xenophobic, and racist speech detection tweets in Greek with an F1 score of 0.97 in their best model.

In this work, we propose a new framework to analyze Polarisation in Online Social Networks. Our approach differentiates from the previous studies as it is based on deep ULMFiT to provide a classification for a large amount of data with high accuracy. We do not use only tweets but also retweets and quotes. Moreover, the approach is more fine-grained as we classify the tweets into five classes (hate speech, offensive, sexism, neutral and positive). Finally, we monitor the evolution of the sentiments mentioned earlier over time in different modern communities in Twitter, such as these with hashtags #Coronavirus, #ClimateChange, #Immigrants, and #MeToo.

III. The Proposed Framework

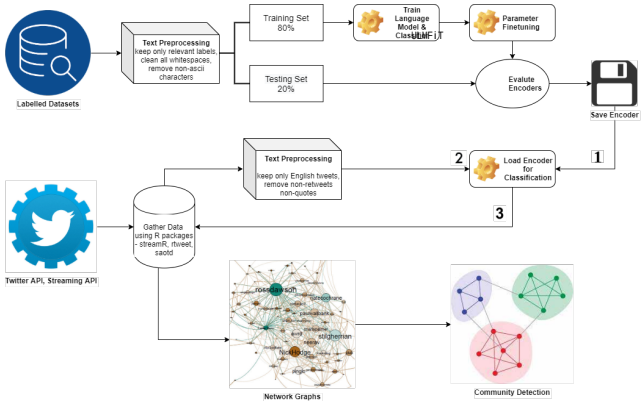

Fig. 1: Process Workflow

The proposed framework is presented in Fig. 1. Labeled data was collected to train a model on multiclass classification tasks, and hashtag-related data was collected from Twitter API to indicate the topics we picked. Text pre-processing involved standardization of white space characters and removing non-ASCII characters on both datasets, while the rest of the tweet text kept its original state. The labeled data was then split with an 80:20 ratio for training and testing a pre-trained language and classifier model. When enough topic-related data had been gathered, it was fed to the classifier model to classify the collected text. Finally, we created network graphs on the classified data, and communities were detected within, based on the sentiment observations. The Twitter data used in this research comes from a collection of tweets with hashtags - **#MeToo, #Immigrants, #ClimateChange** and an addition as of April 2020, **#Coronavirus**. In order to test our framework, we have made use of data from the social media platform Twitter to identify the sentiments expressed towards political topics (and not only) of global importance, such as Immigration, the MeToo movement, and the Coronavirus epidemiology. Topics picked to get Twitter data were based on issues that gained social momentum and were in global news in recent times: **MeToo Movement**, **Immigration**, **Climate Change** and **Coronavirus**. In this context, the tweets related to each topic were classified into one of the following five categories regarding their sentiment polarity: **Hate Speech**, **Offensive**, **Sexism**, **Neutral** and **Positive**. Then, communities were defined through user retweets concerning a topic and revealed Twitter users with influential roles in their respective communities.

A. Labeled Datasets

The lack of reliably labeled datasets puts a large emphasis on available datasets containing human-labeled records to capture the sentiment representation as close as possible. For the **#MeToo** topic, a large Twitter dataset available on *data.world*[2] was used for tweets related to the Me Too Movement[3]. Harvard Dataverse is one of the largest online data repositories available for academic and research purposes. **#ClimateChange** `tweet-id` were collected from one such repository authored by researchers Littman and Wrubel, 2019 [19] from George Washington University, USA. A corpus data was created by combining six labeled datasets and later used to train the model on the multiclass classification task.

Dataset 1: Contains 99,799 annotated `tweet-id` from CrowdFlower users. Used as premise for flagging abusive behavior on Twitter [20],

Dataset 2: In [21] the *Expert* annotators were defined as those having theoretical and applied knowledge of hate speech (recruited among feminist and antiracism activists), while the *Amateur* annotations were obtained from CrowdFlower platform.

Dataset 3: With 25,297 labeled tweet texts, this dataset was also annotated by CrowdFlower users and used in [22].

Dataset 4: Used in [23], Sentiment140[4]. The dataset contains 1.6 million annotated text with emoticons, proving helpful for this paper since the tweets were classified without removing emoticons. Only **positive** sentiment text was used from this dataset, with 800k records.

Dataset 5: Presented by Google Jigsaw during Kaggle's Toxic Comment Classification Challenge [24].

Dataset 6: From 5,500 hand-picked tweets published by Sander's Lab, Sanders Analytics LLC (a now defunct company), the records labelled as **positive** and **neutral** were used from the dataset[5].

B. ULMFiT

Language modeling captures many aspects of language relevant to text classification. Howard and Ruder [25] claim that the language model aims to predict the next word given its previous word. This approach of learning is beneficial in many NLP tasks. Of the noteworthy models that follow language modeling such as Embeddings from Language Models (ELMo) and OpenAI Transformer, Universal Language Model Fine-tuning (ULMFiT), is the *fast.ai*[6] library, which

[2]https://data.world/hamdan/tweets-with-emojis-metoo-2017-10-16
[3]https://en.wikipedia.org/wiki/Me_Too_movement
[4]http://help.sentiment140.com/home
[5]https://github.com/guyz/twitter-sentiment-dataset/blob/master/corpus.csv
[6]https://nlp.fast.ai/classification/2018/05/15/introducing-ulmfit.html

is built on top of *PyTorch*. This allows for straightforward model implementation on NLP tasks. Text classification with ULMFiT is a multi-step method and consists of three main stages, as we can see in Fig. 2:

1) Feeding and training the pre-trained language model on a large text corpus that captures high-level language features. At this stage, the model learns the general features of the language.
2) Fine-tuning the pre-trained language model on target task data. Through transfer learning, the knowledge gained in the source task is shared in the target task. However, the target task data will likely have different features and distribution than the source task data. The language model is then fine-tuned on target task data to reduce learning variance. The model can not only predict the next word but also learns task-specific language features, such as *Twitter mentions, internet slang, emoticons, and hyperlinks*.
3) Fine-tuning the classifier on target task data. The ultimate goal of the model is to provide sentiment classification against fed text, so in a third step, the pre-trained language model is augmented with two linear blocks that give a probability distribution over the target classes (sentiment labels). Sentiment labels applied to datasets in this paper are **hate speech, offensive, sexism, positive, neutral**.

AWD-ULMFiT is based on Long-short term memory networks (LSTMs). LSTMs are a specific recurrent neural network (RNN) architecture, whose design capable of learning long term dependencies using the concept of memory. They were first introduced by [26], then they have been varied and popularised for solving the large diversity of problems. The basic LSTM structure consists of three gates: a forget gate f, an input gate i, which are both made for the cell state's update C_t, and an output gate o that is used to decide the amount of information about the current input x_t that should be saved to the current output cell h_t. The equations of the different gates are as follows:

$$f_t = \sigma(W_f.[h_{t-1}, x_t] + b_f) \tag{1}$$

$$i_t = \sigma(W_i.[h_{t-1}, x_t] + b_i) \tag{2}$$

$$\hat{C}_t = tanh(W_c.[h_{t-1}, x_t] + b_c) \tag{3}$$

$$C_t = f_t \times C_{t-1} + i_t \times \hat{C}_t \tag{4}$$

$$o_t = \sigma(W_o.[h_{t-1}, x_t] + b_o) \tag{5}$$

$$h_t = o_t \times tanh(C_t) \tag{6}$$

ASGD Weight-Dropped LSTM, or AWD-LSTM, is a type of recurrent neural network that employs DropConnect for regularization, as well as NT-ASGD for optimization - non-monotonically triggered averaged SGD - which returns an average of last iterations of weights [27]. AWD-LSTM performs the concept of DropConnect on the hidden-to-hidden weight matrices to avoid overfitting [28].

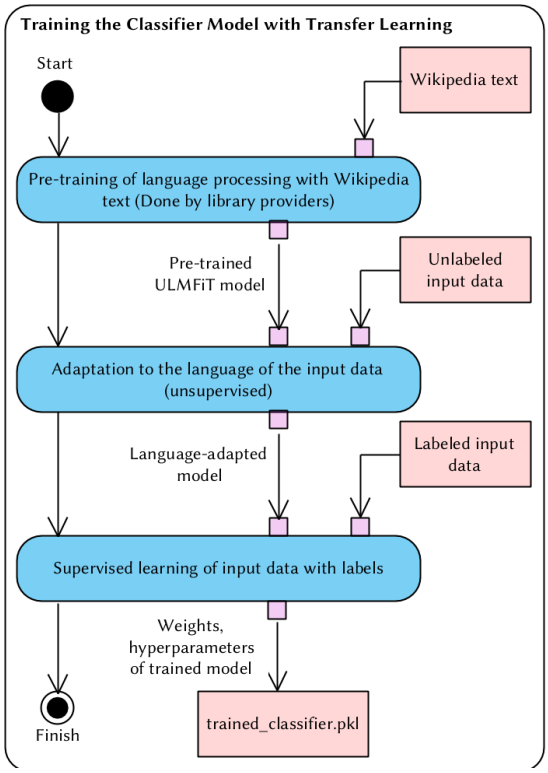

Fig. 2: Training the ULMFiT Classifier

C. Hate Speech Flow Evolution

We analyze topic-based communities of influence within the considered dataset and the evolution of communities over time. The approach is similar to work in [29] on social media data, attuned for our dataset that comprises of topics represented by hashtags **#MeToo, #Immigrants, #ClimateChange** and **#Coronavirus**. Recent findings lead us to understand that the structure of the network we create affects the way in which communities emerge. Cerepnalkoski and Mozetic, 2015 [30] studied Twitter networks based on retweets, providing empirical evidence that the formation of retweet networks reflects a community structure and reveals real-world relationships. The study further stresses that network theory is particularly effective in uncovering hidden structural properties when there is little or no prior real-world knowledge. Furthermore, retweets can be considered a form of influence, thus making the author of a retweet - an influencer. The network structure in this paper is created on retweet relationships, and the nodes represent Twitter accounts on both sides (origin, destination) of a retweet with a directed edge connecting the accounts.
The retweet feature can consist of two parts - retweet and retweet with a comment, better known as quotes. Quotes[7] are retweets with additional comments by the user who is posting a retweet. The additional comment or text is valuable in our case as it considers the possibility of the retweeting user having a polar opposite sentiment of the topic from the sentiment of

[7]https://help.twitter.com/en/using-twitter/how-to-retweet

the user who created the original tweet.

We then try to follow the flow of hate speech and the orientation of sentiments in the evolution of the networks. Thus, we see networks as graphs, and we consider the following graph entities for network creation: i) a node is the Twitter user, ii) a node size is the number of followers, iii) a source edge is a user who created original tweet, iv) a target edge is a user who retweeted original tweet, v) an edge weight is the retweet count.

The created network represented clusters of interactions between social media users of different topics, and these hidden clusters were disclosed using a standard community extraction algorithm [31]. Once the communities were detected in the network, we analyzed the sentiment flow within the community and how it evolved.

Yang and Leskovec, 2012 [32] provided scoring functions that measure the community quality and the connectivity structure of a given set of nodes in a network. For a given set of nodes *N*, function *f(N)* characterizes how community-like is the connectivity of nodes within. Let *G(n,m)* be an undirected graph with *n* nodes, *m* edges and *E(m)* represents the expected number of edges in a random graph. Scoring functions for communities in the network considered are:

- Average degree: $f(N) = \frac{2m}{n}$ is the average degree of the members of *N*.
- Density: $f(N) = \frac{m}{n(n-1)/2}$ is the ratio of actual connections to possible connections that measures the network density of the node set *N*.
- Size: $f(N) = m$ is the number of edges between the members of *N*.
- Modularity: $f(N) = \frac{1}{4}.(m - E(m))$ is $1/4^{th}$ the difference between *m* and *E(m)*.

Modularity is a measure of the structure of a graph, measuring the density of connections within a module or community. Graphs with a high modularity score will have many connections within a community, but only a few pointing outwards to other communities [33]–[36]. We use these scoring above functions to measure the connectedness of communities detected in the network. This information is beneficial to derive inferences about the communities and sentiments they project.

IV. Experimental Results

A. Experimental setups

Our experiments were carried out with HETZNER GPU dedicated server, and specifically GeForce® GTX 1080 (CPU: Intel® Core™ i7-6700 Quad-Core, RAM: 64 GB), with a 3.7 Python version, with 1.0.61 fastai and 1.8.1 PyTorch versions.

B. Language Model and Classifier

The language model is downloaded from *fastai* library, the training data is loaded into the model, and the word occurrences are tokenized. Regarding language model fine-tuning, we specify a minimum word frequency of *1*, meaning the model only tokenizes words with more than one occurrence, with a novel token. Once data in the language model is tokenized, the learner object is used to retrain AWD-LSTM on tokenized data, i.e., target task dataset. Part of the model embedding layers is untrained on initial random weights, i.e., from WikiText-103 (103M tokens) embedding matrix and retraining the entire model risked forgetting existing knowledge within the LSTM. To avoid this, we implement the technique of *freezing* and *unfreezing*, first introduced in [37].

Language model fine-tuning makes a cycle of one learning rate increase phase followed by a decreasing phase. The cycle lasts for one epoch with a maximum learning rate of 0.001 and a momentum between 0.8 and 0.7. Then, we unfreeze all layers and run for 20 epoch cycles in total until the training loss is reduced to an acceptable level. The optimal learning rate is 0.1, and the dropouts are equal to 0.5. The classification used the same structure where the output layer considered the number of classes in the target dataset. There was a gradual unfreeze of layers while running the epochs for model fine-tuning as suggested by Howard and Ruder [38]. The authors propose gradually unfreezing the model starting from the last layer, which contains the slightest general knowledge. Instead of using the same learning rate for all model layers, discriminative fine-tuning allows us to tune each layer with different learning rates. First, we unfreeze the two last layers and fine-tune all unfrozen layers for one epoch. The function slice ($1e^{-4}$, $1e^{-2}$) means that we train every layer with different learning rates ranging from max to min value, and the momentum is between 0.8 and 0.7. Then, we use the unfreeze function, and we run a cycle of 10 epochs and slice($1e^{-5}$, $1e^{-3}$). For training the classifier model, we chose a batch size of 128. The learning rate is 0.1, and the dropouts are 0.5. The model is trained on ten epoch cycles, and the final training accuracy is **85%**.

In Table I, the confusion matrix of the imbalanced test set and the evaluation metrics are presented. The overall precision and recall results were better in the imbalanced dataset in comparison to a balanced dataset. F_{micro} is identical to overall accuracy, i.e., correct predictions divided by all predictions. Since F_{macro} gives every class the same weight and the imbalanced dataset was used, F_{micro} was the more suitable evaluation measure with a value of **85%**. The F_{micro} for the balanced dataset was *82%* and the F_{macro} for the imbalanced and balanced datasets were *83%* and *82%* respectively.

C. Comparison with State of the Art

Next, we test our classifier by using four publicly available datasets from the related work. For the sake of consistency, for each case, we use the same evaluation metrics. The results are presented in Table II. In [22], an overall F_{micro} of 0.91 is found, whereas our classifier gives 0.86 for the same dataset. However, the authors find precision and recall of 0.44 and 0.61 for hate speech, whereas we find 0.49 and 0.59. In [9], the recall for the offensive is 0.91, and we also find 0.91. For the dataset from [39] we find precision and recall for the hate speech 0.64 and 0.81 and no hate speech 0.85 and 0.7 respectively. Moreover, the overall F_{micro} is 0.74. Due to the fact that the authors of the paper [39] use the accuracy of hate speech and no hate speech as well as the overall

TABLE I: *Evaluation Measures on Imbalanced Dataset*

	Actuals								
Predicted	**hate speech**	**neutral**	**offensive**	**positive**	**sexism**	**Total**	**Precision**	**Recall**	**F$_{micro}$**
hate speech	**1118**	100	364	10	8	1600	69.9%	63.7%	67%
neutral	125	**3430**	191	80	48	3874	88.5%	85.6%	87%
offensive	471	172	**3341**	45	9	4038	82.7%	84.2%	83%
positive	31	245	66	**2884**	21	3247	88.8%	95.4%	92%
sexism	9	61	4	3	**550**	627	87.7%	86.5%	87%
Total	1754	4008	3966	3022	636	13386	**84.6%**	**84.6%**	**85%**

TABLE II: *Evaluation results with other datasets*

	Sentiment	**Precision**	**Recall**	**F$_{micro}$**
Dataset 1 [22]	Hate Speech	0.49	0.59	0.86
Dataset 1 [22]	No Hate Speech	0.4	0.87	0.86
Dataset 2 [39]	Hate Speech	0.64	0.81	0.74
Dataset 2 [39]	No Hate Speech	0.85	0.7	0.74
Dataset 3 [40]	Hate Speech	0.70	0.60	0.62
Dataset 3 [40]	No Hate Speech	0.54	0.7	0.62
Dataset 4 [9]	Hate Speech	0.59	0.65	0.87
Dataset 4 [9]	No Hate Speech	0.93	0.91	0.87
Dataset 5 [20]	Hate Speech	0.45	0.59	0.79
Dataset 5 [20]	No Hate Speech	0.93	0.75	0.81

TABLE III: *Number of tweets for all the topics*

	Tweets	**Retweets/Quotes**	**Without neutral**
Immigrants	17,824	12,734	4,170
MeToo	196,749	131,477	27,608
ClimateChange	879,692	713,348	54,486
Coronavirus	134,129	109,765	9,156

accuracy, we compute the three accuracy for our approach. As a result, we achieve a 0.88 for the hate speech accuracy that outperforms the 0.76 from the best model of [39]. Our no-hate speech accuracy is 0.7, whereas Gibert et al. score 0.8 for the best model. Our overall accuracy is 0.79 that is slightly better than 0.78. The third dataset is from [40] and the overall F_{macro} of the best-performed team is 0.65 and from the second best-performed team 0.57. We find an F1 score of 0.62. In particular, we achieve 0.7 and 0.6 for precision and recall for the hate speech and 0.54 and 0.64 for the no hate speech. For the dataset from [9], our results show precision and recall of 0.59 and 0.65 for hate speech and 0.93 and 0.92 for no hate speech, whereas the overall F_{micro} is 0.87. Finally, with the dataset from [20], we achieve precision and recall of 0.45 and 0.59 for hate speech, 0.88 and 0.85 for the offensive, and 0.75 and 0.81 for the normal. The overall F_{micro} is 0.79.

D. Case study - Sentiment Orientation

We evaluate the quality of the results of our approach in a way that it does involve internal score measurements (number of tweets, average degree, density, network size, modularity) and the sentiment orientation among users and communities during the evolution of the network. We note that for all four networks, we kept neutral sentiment for realistic visualization of the flow of sentiment, and it is predominant in comparison with other sentiments such as hate speech. When we describe the most popular sentiment per snapshot, we exclude the neutral and refer to the next one in the order. Table III, reports the number of tweets and retweets, including quotes for the four topics. Due to space limitations, we only report the number of tweets per month for the #MeToo in Fig. 3.

In order to follow the sentiment orientation during the evolution of the network, we split our networks into different snapshots, and for each snapshot, we use Louvain [31] to detect the communities. Louvain algorithm maximizes modularity score for each community, where the modularity quantifies the quality of an assignment of nodes to communities. Thus, we evaluate how densely the nodes are connected within a community, compared to how connected they would be in a random network. This evaluation will also provide insights into the correlation between network density and sentiment. For example, communities immersed in hate speech could be closely linked because there is a constant controversy among users [41]. The resulting communities and the sentiment orientation of Coronavirus are depicted in Fig. 4.

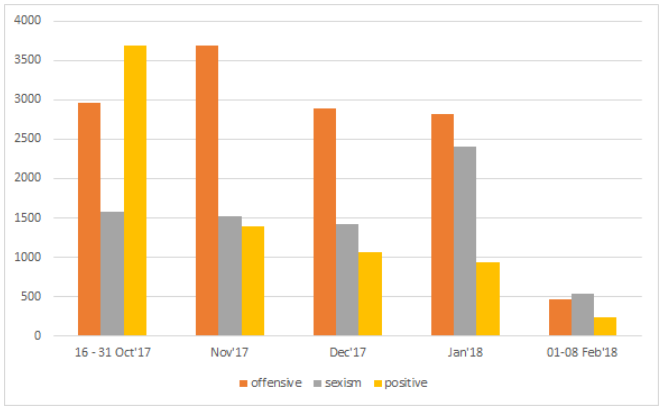

Fig. 3: Number of tweets per month for #MeToo

#Coronavirus: We analyzed the tweets as the lockdown period began in March 2020 across many countries simultaneously. In this context, we use a publicly available dataset [42]. The initial tweet texts were classified with more **positive** sentiments, and communities gradually began to show more **hate speech** sentiments. As expected, the size of communities increased from 604 to 2,731 (with the Label Propagation algorithm) in the period from March 2020 to April 2020, where we had the peak in discussions about the imposed measures for fighting the coronavirus. The average degree per community during the evolution ranges between 5 and 6, and the modularity is high in all snapshots, which implies interactions (tweets/retweets) between users of the same communities. The shift in popular sentiment is visible from the graphs in Fig. 4, as it moves

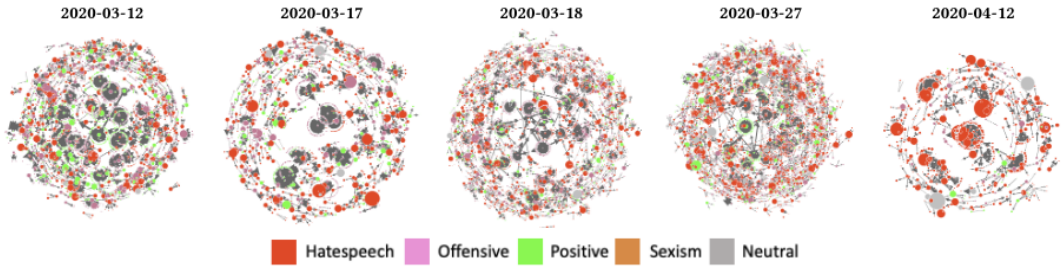

Fig. 4: #**Coronavirus** sentiment orientation over time. Modularity: [0.94, 0.77, 0.96, 0.89]

from more tweets getting classified as **positive** to **hate speech**. In addition, we observe community crossovers that carry forward the sentiment as the graphs present a notable flow in sentiments through community evolution over time (Fig. 4).

#ClimateChange: This was the most extensive dataset with 879,000 samples for 43 days, and 16% (143,000) of the classified dataset was used for network analysis and further reduced to only retweets. The dataset produced 20 graphs, one for each day, and the networks had close representations on most days. The communities grew and became increasingly dense over time, implying several tweet-retweet interactions. We identified 1,656 communities in `2018-08-13` that increase significantly in 2018-08-21 to 5,043 (with the Weekly Connected Components algorithm). Moreover, the average size increased from 14.82 to 52.94, meaning 15 to 53 Twitter users per community. Multiple crossovers were observed however overall sentiment of the communities detected is **hate speech**. We also observed a node that stood out in four of the network graphs and based on the multiple presence and node size, i.e., followers count, it was detected as an influencer of sentiments in its community.

#MeToo: In the first snapshot (`November 2017`) we see a high sentiment polarity where members within the same community represent multiple sentiments **sexism, offensive** and **positive**. The change in the size of communities follows the same trends as in the previous cases. Modularity in these snapshots is too high, implying how sensitive the topic is and how much it kept people from discussing and arguing within the same community. A study[8] conducted by Pew Research Center shows that **#MeToo** was used 19 million times since the initial tweet due to the presence of many celebrities on Twitter. We also observe an outflow of emotions from one community to another, possibly due to intersecting nodes. In addition, the nodes in the center of each snapshot are Twitter users who created the original tweet, which several other users later retweeted. In time instance `December 2017`, the nodes at the center of the graph appear prominent, indicating high followers count on Twitter. Likewise, the community members who retweet seem to reflect their sentiments. Communities that retweeted grew in size from `2017-11` to `2018-01` with the average size of communities, i.e., number of nodes present in each community to range between 5.48 and 7.02 (6–7 Twitter users).

#Immigrants: The sentiment of **hate speech** can be seen more in higher prominence as the time instances pass, with more significant node sizes present in the center of the community, detecting possible influencers. Although we observe births and deaths of communities during the evolution of this network, the flow of **hate speech** remains stable over time by flowing through the tweet-retweet interactions. Finally, the average size ranges from 3.69 to 7.12 (4-7 Twitter users), with the size of communities to follow the same trend as in the previous cases. Due to space limitations, we only report the summary statistics on the two large datasets, i.e., Coronavirus and Climate Change in Tables IVa and IVb. We refer to the Weakly Connected Components as WCC, the Label Propagation as LP, and the Asynchronous Label Propagation as ALP.

Time Instance	Algorithm	# of Communities	Avg. Size	Avg. Degree	Popular Sentiment
2020-03-12	WCC	589	5.24	1.18	Positive
2020-03-17	WCC	939	4.87	1.18	Positive
2020-03-18	WCC	1,609	5.77	1.18	Hate speech
2020-03-27	WCC	2,489	5.86	1.17	Hate speech
2020-04-12	WCC	2,627	5.89	1.17	Hate speech
2020-03-12	LP	604	5.1	2.38	Positive
2020-03-17	LP	965	4.74	2.37	Positive
2020-03-18	LP	1,719	5.4	2.39	Hate speech
2020-03-27	LP	2,731	5.34	2.39	Hate speech
2020-04-12	LP	2,731	5.36	2.4	Hate speech
2020-03-12	ALP	605	5.1	2.38	Positive
2020-03-17	ALP	964	4.75	2.37	Positive
2020-03-18	ALP	1,715	5.41	2.39	Hate speech
2020-03-27	ALP	2,719	5.36	2.4	Hate speech
2020-04-12	ALP	2880	5.37	2.4	Hate speech

(a)

Time Instance	Algorithm	# of Communities	Avg. Size	Avg. Degree	Popular Sentiment
2018-08-13	WCC	1,656	14.82	1.16	Hate speech
2018-08-14	WCC	2,629	19.71	1.17	Hate speech
2018-08-15	WCC	3,310	22.54	1.17	Hate speech
2018-08-17	WCC	4,196	26.08	1.16	Hate speech
2018-08-21	WCC	5043	52.94	1.14	Hate speech
2018-08-13	LP	3,059	8.02	2.52	Hate speech
2018-08-14	LP	5,689	9.1	2.55	Hate speech
2018-08-15	LP	8,113	9.18	2.56	Hate speech
2018-08-17	LP	11,459	9.54	2.56	Hate speech
2018-08-21	LP	14,717	18.12	2.54	Hate speech
2018-08-13	ALP	2,975	8.24	2.55	Hate speech
2018-08-14	ALP	5,579	9.28	2.59	Hate speech
2018-08-15	ALP	7,878	9.46	2.61	Hate speech
2018-08-17	ALP	10,940	9.99	2.61	Hate speech
2018-08-21	ALP	15,866	16.81	2.62	Hate speech

(b)

TABLE IV: (a) **#Coronavirus** communities: March - April 2020, and (b) **#Climate Change** communities: 13 – 21 August 2018

V. Conclusions

In this paper, we propose a deep learning-based framework that collects tweets related to topics such as coronavirus pandemic, climate change, immigrants, and MeToo movement and

[8]https://www.pewresearch.org/fact-tank/2018/10/11/how-social-media-users-have-discussed-sexual-harassment-since-metoo-went-viral/

classifies them into five classes: i) Hate Speech, ii) Offensive iii) Sexism iv) Neutral and v) Positive. Universal Language Model Fine-tuning (ULMFiT) is used for the five-class classification. Then, we detect communities within the networks of these topics, and we monitor the sentiment orientation during the evolution of the networks. The experimental results show that the proposed framework outperforms the baseline models regarding the accuracy and F-score evaluation metrics. The proposed approach is also tested on other public datasets and both on multiclass classification and binary classification. We plan to apply our framework to more data from more prolonged periods in the future. For instance, we would like to use our trained classifier with the data from **#Coronavirus** from April 2020 until today and detect how and why the sentiment changed within the different communities.

REFERENCES

[1] J. Bollen, A. Pepe, and H. Mao, "Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena," *CORR*, 2009.

[2] L. Belkin, T. Kurtzberg, and C. Naquin, "Emotional contagion in the online environment: Investigating the dynamics and implications of emotional encounters in mixed-motive situations in the electronic context," *SSRN Electronic Journal*, 07 2006.

[3] R. Hillmann and M. Trier, "Sentiment polarization and balance among users in online social networks," in *AMCIS*, vol. 24, 2012.

[4] K. Seo, R. Pan, and A. Panasyuk, "Detecting communities by sentiment analysis of controversial topics," in *SBP-BRiMS*, vol. 9708, Germany, 2016, pp. 206–215.

[5] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," 2011.

[6] S. Frenda, B. Ghanem, M. Montes, and P. Rosso, "Online hate speech against women: Automatic identification of misogyny and sexism on twitter," *Journal of Intelligent & Fuzzy Systems*, vol. 36, pp. 4743–4752, 2019.

[7] S. Sharma, S. Agrawal, and M. Shrivastava, "Degree based classification of harmful speech using twitter data," in *TRAC*, Santa Fe, New Mexico, USA, 2018, pp. 106–112.

[8] B. Mathew, R. Dutt, P. Goyal, and A. Mukherjee, "Spread of hate speech in online social media," 2018.

[9] E. Wulczyn, N. Thain, and L. Dixon, "Ex machina: Personal attacks seen at scale," 2016.

[10] P. Tsantarliotis, E. Pitoura, and P. Tsaparas, "Troll vulnerability in online social networks," in *ASONAM*, 2016, pp. 1394–1396.

[11] P. Fornacciari, M. Mordonini, and M. Tomaiuolo, "Social network and sentiment analysis on twitter: Towards a combined approach." 01 2015, pp. 53–64.

[12] Z. Xu and Q. Yang, "Analyzing user retweet behavior on twitter," in *ASONAM*, 08 2012, pp. 46–50.

[13] L. Jiang, M. Yu, M. Zhou, X. Liu, and T. Zhao, "Target-dependent twitter sentiment classification." 01 2011, pp. 151–160.

[14] M. Speriosu, N. Sudan, S. Upadhyay, and J. Baldridge, "Twitter polarity classification with label propagation over lexical links and the follower graph," 07 2011, pp. 53–63.

[15] A. M. Founta, D. Chatzakou, N. Kourtellis, J. Blackburn, A. Vakali, and I. Leontiadis, "A unified deep learning architecture for abuse detection," in *WebSci*, New York, NY, USA, 2019, p. 105–114.

[16] R. Cao, R. K.-W. Lee, and T.-A. Hoang, "Deephate: Hate speech detection via multi-faceted text representations," in *12th ACM Conference on Web Science*, ser. WebSci '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 11–20. [Online]. Available: https://doi.org/10.1145/3394231.3397890

[17] M. Rizoiu, T. Wang, G. Ferraro, and H. Suominen, "Transfer learning for hate speech detection in social media," *CoRR*, vol. abs/1906.03829, 2019. [Online]. Available: http://arxiv.org/abs/1906.03829

[18] K. Perifanos and D. Goutsos, "Multimodal hate speech detection in greek social media," *Multimodal Technologies and Interaction*, vol. 5, no. 7, 2021. [Online]. Available: https://www.mdpi.com/2414-4088/5/7/34

[19] J. Littman and L. Wrubel, "Climate Change Tweets Ids," 2019. [Online]. Available: https://doi.org/10.7910/DVN/5QCCUU

[20] A.-M. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, and N. Kourtellis, "Large scale crowdsourcing and characterization of twitter abusive behavior," in *ICWSM*, 2018.

[21] Z. Waseem and D. Hovy, "Hateful symbols or hateful people? predictive features for hate speech detection on twitter," in *NAACL Student Research Workshop*, San Diego, California, June 2016, pp. 88–93.

[22] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language," 2017.

[23] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *Processing*, vol. 150, 01 2009.

[24] B. van Aken, J. R., R. Krestel, and A. Löser, "Challenges for toxic comment classification: An in-depth error analysis," 2018.

[25] S. Ruder, "NLP's ImageNet moment has arrived," https://ruder.io/nlp-imagenet/, 2018.

[26] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, p. 1735–1780, nov 1997.

[27] S. Merity, N. S. Keskar, and R. Socher, "Regularizing and Optimizing LSTM Language Models," *arXiv preprint arXiv:1708.02182*, 2017.

[28] L. Wan, M. Zeiler, S. Zhang, Y. Le Cun, and R. Fergus, "Regularization of neural networks using dropconnect," in *Proceedings of the 30th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, S. Dasgupta and D. McAllester, Eds., vol. 28, no. 3. Atlanta, Georgia, USA: PMLR, 17–19 Jun 2013, pp. 1058–1066. [Online]. Available: https://proceedings.mlr.press/v28/wan13.html

[29] M. A. Smith, L. Rainie, B. Shneiderman, and I. Himelboim, "Mapping twitter topic networks: From polarized crowds to community clusters," 2014. [Online]. Available: https://www.pewresearch.org/internet/2014/02/20/mapping-twitter-topic-networks-from-polarized-crowds-to-community-clusters/

[30] D. Cerepnalkoski and I. Mozetic, "A retweet network analysis of the European Parliament," in *SITIS*. IEEE, 2015, pp. 350–357.

[31] H. Lu, M. Halappanavar, and A. Kalyanaraman, "Parallel heuristics for scalable community detection," *Parallel Comput.*, vol. 47, pp. 19–37, 2015.

[32] J. Yang and J. Leskovec, "Defining and evaluating network communities based on ground-truth," 2012.

[33] M. Girvan and M. E. J. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, p. 7821–7826, Jun 2002. [Online]. Available: http://dx.doi.org/10.1073/pnas.122653799

[34] G. Agarwal and D. Kempe, "Modularity-maximizing graph communities via mathematical programming," *The European Physical Journal B*, vol. 66, no. 3, p. 409–418, Nov 2008. [Online]. Available: http://dx.doi.org/10.1140/epjb/e2008-00425-1

[35] M. E. J. Newman, "Fast algorithm for detecting community structure in networks," *Physical Review E*, vol. 69, no. 6, Jun 2004. [Online]. Available: http://dx.doi.org/10.1103/PhysRevE.69.066133

[36] U. Brandes, D. Delling, M. Gaertler, R. Gorke, M. Hoefer, Z. Nikoloski, and D. Wagner, "On modularity clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 20, no. 2, pp. 172–188, 2008.

[37] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *EMNLP*, Copenhagen, Denmark, 2017, pp. 1615–1625.

[38] J. Howard and S. Ruder, "Universal language model fine-tuning for text classification," 2018.

[39] O. de Gibert, N. Perez, A. García-Pablos, and M. Cuadros, "Hate speech dataset from a white supremacy forum," in *ALW2*, Brussels, Belgium, Oct. 2018, pp. 11–20.

[40] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel P., P. Rosso, and M. Sanguinetti, "SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter," in *SemEval@NAACL-HLT*, Minneapolis, Minnesota, USA, 2019, pp. 54–63.

[41] A. Matakos, E. Terzi, and P. Tsaparas, "Measuring and moderating opinion polarization in social networks," *Data Min. Knowl. Discov.*, vol. 31, no. 5, pp. 1480–1505, 2017.

[42] R. Lamsal, "Coronavirus (covid-19) tweets dataset," 2020. [Online]. Available: https://dx.doi.org/10.21227/781w-ef42