Multilingual and Multi-Aspect Hate Speech Analysis

Abstract

Current research on hate speech analysis is typically oriented towards monolingual and single classification tasks. In this paper, we present a new multilingual multi-aspect hate speech analysis dataset and use it to test the current state-of-the-art multilingual multitask learning approaches. We evaluate our dataset in various classification settings, then we discuss how to leverage our annotations in order to improve hate speech detection and classification in general.

Introduction

With the expanding amount of text data generated on different social media platforms, current filters are insufficient to prevent the spread of hate speech. Most internet users involved in a study conducted by the Pew Research Center report having been subjected to offensive name calling online or witnessed someone being physically threatened or harassed online. Additionally, Amnesty International within Element AI have lately reported that many women politicians and journalists are assaulted every 30 seconds on Twitter. This is despite the Twitter policy condemning the promotion of violence against people on the basis of race, ethnicity, national origin, sexual orientation, gender identity, religious affiliation, age, disability, or serious disease. Hate speech may not represent the general opinion, yet it promotes the dehumanization of people who are typically from minority groups BIBREF0, BIBREF1 and can incite hate crime BIBREF2.

Moreover, although people of various linguistic backgrounds are exposed to hate speech BIBREF3, BIBREF2, English is still at the center of existing work on toxic language analysis. Recently, some research studies have been conducted on languages such as German BIBREF4, Arabic BIBREF5, and

Italian BIBREF6. However, such studies usually use monolingual corpora and do not contrast, or examine the correlations between online hate speech in different languages. On the other hand, tasks involving more than one language such as the hatEval task, which covers English and Spanish, include only separate classification tasks, namely (a) women and immigrants as target groups, (b) individual or generic hate and, (c) aggressive or non-aggressive hate speech.

Treating hate speech classification as a binary task may not be enough to inspect the motivation and the behavior of the users promoting it and, how people would react to it. For instance, the hateful tweets presented in Figure FIGREF5 show toxicity directed towards different targets, with or without using slurs, and generating several types of reactions. We believe that, in order to balance between truth and subjectivity, there are at least five important aspects in hate speech analysis. Hence, our annotations indicate (a) whether the text is direct or indirect; (b) if it is offensive, disrespectful, hateful, fearful out of ignorance, abusive, or normal; (c) the attribute based on which it discriminates against an individual or a group of people; (d) the name of this group; and (e) how the annotators feel about its content within a range of negative to neutral sentiments. To the best of our knowledge there are no other hate speech datasets that attempt to capture fear out of ignorance in hateful tweets or examine how people react to hate speech. We claim that our multi-aspect annotation schema would provide a valuable insight into several linguistic and cultural differences and bias in hate speech.

We use Amazon Mechanical Turk to label around 13,000 potentially derogatory tweets in English, French, and Arabic based on the above mentioned aspects and, regard each aspect as a prediction task. Since in natural language processing, there is a peculiar interest in multitask learning, where different tasks can be used to help each other BIBREF7, BIBREF8, BIBREF9, we use a unified model to handle the annotated data in all three languages and five tasks. We adopt BIBREF8 as a learning algorithm adapted to loosely related tasks such as our five annotated aspects and, use the Babylon cross-lingual embeddings BIBREF10 to align the three languages. We compare the multilingual multitask learning settings with

monolingual multitask, multilingual single-task, and monolingual single-task learning settings respectively. Then, we report the performance results of the different settings and discuss how each task affects the remaining ones. We release our dataset and code to the community to extend research work on multilingual hate speech detection and classification.

Related Work

There is little consensus on the difference between profanity and hate speech and, how to define the latter BIBREF17. As shown in Figure FIGREF11, slurs are not an unequivocal indicator of hate speech and can be part of a non-aggressive conversation, while some of the most offensive comments may come in the form of subtle metaphors or sarcasm BIBREF18. Consequently, there is no existing human annotated vocabulary that explicitly reveals the presence of hate speech, which makes the available hate speech corpora sparse and noisy BIBREF19.

Given the subjectivity and the complexity of such data, annotation schemes have rarely been made fine-grained. Table TABREF10 compares different labelsets that exist in the literature. For instance, BIBREF12 use racist, sexist, and normal as labels; BIBREF13 label their data as hateful, offensive (but not hateful), and neither, while BIBREF16 present an English dataset that records the target category based on which hate speech discriminates against people, such as ethnicity, gender, or sexual orientation and ask human annotators to classify the tweets as hate and non hate. BIBREF15 label their data as offensive, abusive, hateful, aggressive, cyberbullying, spam, and normal. On the other hand, BIBREF20 have chosen to detect ideologies of hate speech counting 40 different hate ideologies among 13 extremist hate groups.

The detection of hate speech targets is yet another challenging aspect of the annotation. BIBREF21 report the bias that exists in the current datasets towards identity words, such as women, which may later

cause false predictions. They propose to debias gender identity word embeddings with additional data for training and tuning their binary classifier. We address this false positive bias problem and the common ambiguity of target detection by asking the annotators to label target attributes such as origin, gender, or religious affiliation within 16 named target groups such as refugees, or immigrants.

Furthermore, BIBREF22 have reproduced the experiment of BIBREF12 in order to study how hate speech affects the popularity of a tweet, but discovered that some tweets have been deleted. For replication purposes, we provide the community with anonymized tweet texts rather than IDs.

Non-English hate speech datasets include Italian, German, Dutch, and Arabic corpora. BIBREF6 present a dataset of Italian tweets, in which the annotations capture the degree of intensity of offensive and aggressive tweets, in addition to whether the tweets are ironic and contain stereotypes or not. BIBREF2 have collected more than 500 German tweets against refugees, and annotated them as hateful and not hateful. BIBREF23 detect bullies and victims among youngsters in Dutch comments on AskFM, and classify cyberbullying comments as insults or threats. Moreover, BIBREF5 provide a corpus of Arabic sectarian speech.

Another predominant phenomenon in hate speech corpora is code switching. BIBREF24 present a dataset of code mixed Hindi-English tweets, while BIBREF25 report the presence of Hindi tokens in English data and use multilingual word embeddings to deal with this issue when detecting toxicity. Similarly, we use such embeddings to take advantage of the multilinguality and comparability of our corpora during the classification.

Our dataset is the first trilingual dataset comprising English, French, and Arabic tweets that encompasses various targets and hostility types. Additionally, to the best of our knowledge, this is the first work that examines how annotators react to hate speech comments.

To fully exploit the collected annotations, we tested multitask learning on our dataset. Multitask learning BIBREF7 allows neural networks to share parameters with one another and, thus, learn from related tasks. It has been used in different NLP tasks such as parsing BIBREF9, dependency parsing BIBREF26, neural machine translation BIBREF27, sentiment analysis BIBREF28, and other tasks. Multitask learning architectures tackle challenges that include sharing the label space and the question of private and shared space for loosely related tasks BIBREF8, for which techniques may involve a massive space of potential parameter sharing architectures.

Dataset

In this section, we present our data collection methodology and annotation process.

Dataset ::: Data Collection

Considering the cultural differences and commonly debated topics in the main geographic regions where English, French, and Arabic are spoken, searching for equivalent terms in the three languages led to different results at first. Therefore, after looking for 1,000 tweets per 15 more or less equivalent phrases in the three languages, we revised our search words three times by questioning the results, adding phrases, and taking off unlikely ones in each of the languages. In fact, we started our data collection by searching for common slurs and demeaning expressions such as “go back to where you come from”. Then, we observed that discussions about controversial topics, such as feminism in general, illegal immigrants in English, Islamo-gauchisme (“Islamic leftism") in French, or Iran in Arabic were more likely to provoke disputes, comments filled with toxicity and thus, notable insult patterns that we looked for in subsequent search rounds.

Dataset ::: Linguistic Challenges

All of the annotated tweets include original tweets only, whose content has been processed by (1) deleting unarguably detectable spam tweets, (2) removing unreadable characters and emojis, and (3) masking the names of mentioned users using @user and potentially enclosed URLs using @url. As a result, annotators had to face the lack of context generated by this normalization process.

Furthermore, we perceived code-switching in English where Hindi, Spanish, and French tokens appear in the tweets. Some French tweets also contain Romanized dialectal Arabic tokens generated by, most likely, bilingual North African Twitter users. Hence, although we eliminated most of these tweets in order to avoid misleading the annotators, the possibly remaining ones still added noise to the data.

One more challenge that the annotators and ourselves had to tackle, consisted of Arabic diglossia and switching between different Arabic dialects and Modern Standard Arabic (MSA). While MSA represents the standardized and literary variety of Arabic, there are several Arabic dialects spoken in North Africa and the Middle East in use on Twitter. Therefore, we searched for derogatory terms adapted to different circumstances, and acquired an Arabic corpus that combines tweets written in MSA and Arabic dialects. For instance, the tweet shown in Figure FIGREF5 contains a dialectal slur that means “maiden.”

Dataset ::: Annotation Process

We rely on the general public opinion and common linguistic knowledge to assess how people view and react to hate speech. Given the subjectivity and difficulty of the task, we reminded the annotators not to let their personal opinions about the topics being discussed in the tweets influence their annotation decisions.

Our annotation guidelines explained the fact that offensive comments and hate do not necessarily come in the form of profanity. Since different degrees of discrimination work on the dehumanization of

individuals or groups of people in distinct ways, we chose not to annotate the tweets within two or three classes. For instance, a sexist comment can be disrespectful, hateful, or offensive towards women. Our initial labelset was established in conformity with the prevalent anti-social behaviors people tend to deal with. We also chose to address the problem of false positives caused by the misleading use of identity words by asking the annotators to label both the target attributes and groups.

Dataset ::: Annotation Process ::: Avoiding scams

To prevent scams, we also prepared three annotation guideline forms and three aligned labelsets written in English, French, and Modern Standard Arabic with respect to the language of the tweets to be annotated.

We requested native speakers to annotate the data and chose annotators with good reputation scores (more than 0.90). We informed the annotator in the guidelines, that in case of noticeable patterns of random labeling on a substantial number of tweets, their work will be rejected and we may have to block them. Since the rejection affects the reputation of the annotators and their chances to get new tasks on Amazon Mechanical Turk, well-reputed annotators are usually reliable. We have divided our corpora into smaller batches on Amazon Mechanical Turk in order to facilitate the analysis of the annotations of the workers and, fairly identify any incoherence patterns possibly caused by the use of an automatic translation system on the tweets, or the repetition of the same annotation schema. If we reject the work of a scam, we notify them, then reassign the tasks to other annotators.

Dataset ::: Pilot Dataset

We initially put samples of 100 tweets in each of the three languages on Amazon Mechanical Turk. We showed the annotators the tweet along with lists of labels describing (a) whether it is direct or indirect

hate speech; (b) if the tweet is dangerous, offensive, hateful, disrespectful, confident or supported by some URL, fearful out of ignorance, or other; (c) the target attribute based on which it discriminates against people, specifically, race, ethnicity, nationality, gender, gender identity, sexual orientation, religious affiliation, disability, and other ("other" could refer to political ideologies or social classes.); (d) the name of its target group, and (e) whether the annotators feel anger, sadness, fear or nothing about the tweets.

Each tweet has been labeled by three annotators. We have provided them with additional text fields to fill in with labels or adjectives that would (1) better describe the tweet, (2) describe how they feel about it more accurately, and (3) name the group of people the tweet shows bias against. We kept the most commonly used labels from our initial labelset, took off some of the initial class names and added frequently introduced labels, especially the emotions of the annotators when reading the tweets and the names of the target groups. For instance, after this step, we have ended up merging race, ethnicity, nationality into one label origin given common confusions we noticed and; added disgust and shock to the emotion labelset; and introduced socialists as a target group label since many annotators have suggested these labels.

Dataset ::: Final Dataset

The final dataset is composed of a pilot corpus of 100 tweets per language, and comparable corpora of 5,647 English tweets, 4,014 French tweets, and 3,353 Arabic tweets. Each of the annotated aspects represents a classification task of its own, that could either be evaluated independently, or, as intended in this paper, tested on how it impacts other tasks. The different labels are designed to facilitate the study of the correlations between the explicitness of the tweet, the type of hostility it conveys, its target attribute, the group it dehumanizes, how different people react to it, and the performance of multitask learning on the five tasks. We assigned each tweet to five annotators, then applied majority voting to each of the

labeling tasks. Given the numbers of annotators and labels in each annotation sub-task, we allowed multilabel annotations in the most subjective classification tasks, namely the hostility type and the annotator's sentiment labels, in order to keep the right human-like approximations. If there are two annotators agreeing on two labels respectively, we add both labels to the annotation.

The average Krippendorff scores for inter-annotator agreement (IAA) are 0.153, 0.244, and 0.202 for English, French, and Arabic respectively, which are comparable to existing complex annotations BIBREF6 given the nature of the labeling tasks and the number of labels.

We present the labelset the annotators refer to, and statistics of our annotated data in the following.

Dataset ::: Final Dataset ::: Directness label

Annotators determine the explicitness of the tweet by labeling it as direct or indirect speech. This should be based on whether the target is explicitly named, or less easily discernible, especially if the tweet contains humor, metaphor, or figurative speech. Table TABREF20 shows that even when partly using equivalent keywords to search for candidate tweets, there are still significant differences in the resulting data.

Dataset ::: Final Dataset ::: Hostility type

To identify the hostility type of the tweet, we stick to the following conventions: (1) if the tweet sounds dangerous, it should be labeled as abusive; (2) according to the degree to which it spreads hate and the tone its author uses, it can be hateful, offensive or disrespectful; (3) if the tweet expresses or spreads fear out of ignorance against a group of individuals, it should be labeled as fearful; (4) otherwise it should be annotated as normal. We define this task to be multilabel. Table TABREF20 shows that hostility types are

relatively consistent across different languages and offensive is the most frequent label.

Dataset ::: Final Dataset ::: Target attribute

After annotating the pilot dataset, we noticed common misconceptions regarding race, ethnicity, and nationality, therefore we merged these attributes into one label origin. Then, we asked the annotators to determine whether the tweet insults or discriminates against people based on their (1) origin, (2) religious affiliation, (3) gender, (4) sexual orientation, (5) special needs or (6) other. Table TABREF20 shows there are fewer tweets targeting disability in Arabic compared to English and French and no tweets insulting people based on their sexual orientation which may be due to the fact that the labels of gender, gender identity, and sexual orientation use almost the same wording. On the other hand, French contains a small number of tweets targeting people based on their gender in comparison to English and Arabic. We have observed significant differences in terms of target attributes in the three languages. More data may help us examine the problems affecting targets of different linguistic backgrounds.

Dataset ::: Final Dataset ::: Target group

We determined 16 common target groups tagged by the annotators after the first annotation step. The annotators had to decide on whether the tweet is aimed at women, people of African descent, Hispanics, gay people, Asians, Arabs, immigrants in general, refugees; people of different religious affiliations such as Hindu, Christian, Jewish people, and Muslims; or from political ideologies socialists, and others. We also provided the annotators with a category to cover hate directed towards one individual, which cannot be generalized. In case the tweet targets more than one group of people, the annotators should choose the group which would be the most affected by it according to them. Table TABREF10 shows the counts of the five categories out of 16 that commonly occur in the three languages. In fact, most of the tweets target individuals or fall into the “other” category. In the latter case, they may target people with different

political views such as liberals or conservatives in English and French, or specific ethnic groups such as Kurdish people in Arabic. English tweets tend to have more tweets targeting people with special needs, due to common language-specific demeaning terms used in conversations where people insult one another. Arabic tweets contain more hateful comments towards women for the same reason. On the other hand, the French corpus contains more tweets that are offensive towards African people, due to hateful comments generated by debates about immigrants.

Dataset ::: Final Dataset ::: Sentiment of the annotator

We claim that the choice of a suitable emotion representation model is key to this sub-task, given the subjective nature and social ground of the annotator's sentiment analysis. After collecting the annotation results of the pilot dataset regarding how people feel about the tweets, and observing the added categories, we adopted a range of sentiments that are in the negative and neutral scales of the hourglass of emotions introduced by BIBREF29. This model includes sentiments that are connected to objectively assessed natural language opinions, and excludes what is known as self-conscious or moral emotions such as shame and guilt. Our labels include shock, sadness, disgust, anger, fear, confusion in case of ambivalence, and indifference. This is the second multilabel task of our model.

Table TABREF20 shows more tweets making the annotators feel disgusted and angry in English, while annotators show more indifference in both French and Arabic. A relatively more frequent label in both French and Arabic is shock, therefore reflecting what some of the annotators were feeling during the labeling process.

Experiments

We report and discuss the results of five classification tasks: (1) the directness of the speech, (2) the

hostility type of the tweet, (3) the discriminating target attribute, (4) the target group, and (5) the annotator's sentiment.

Experiments ::: Models

We compare both traditional baselines using bag-of-words (BOW) as features on Logistic regression (LR), and deep learning based methods.

For deep learning based models, we run bidirectional LSTM (biLSTM) models with one hidden layer on each of the classification tasks. Deeper BiLSTM models performed poorly due to the size of the tweets. We chose to use Sluice networks BIBREF8 since they are suitable for loosely related tasks such as the annotated aspects of our corpora.

We test different models, namely single task single language (STSL), single task multilingual (STML), and multitask multilingual models (MTML) on our dataset. In multilingual settings, we tested Babylon multilingual word embeddings BIBREF10 and MUSE BIBREF30 on the different tasks. We use Babylon embeddings since they appear to outperform MUSE on our data.

Sluice networks BIBREF8 learn the weights of the neural networks sharing parameters (sluices) jointly with the rest of the model and share an embedding layer, Babylon embeddings in our case, that associates the elements of an input sequence. We use a standard 1-layer BiLSTM partitioned into two subspaces, a shared subspace and a private one, forced to be orthogonal through a regularization penalty term in the loss function in order to enable the multitask network to learn both task-specific and shared representations. The hidden layer has a dimension of 200, the learning rate is initially set to 0.1 with a learning rate decay, and we use the DyNet BIBREF31 automatic minibatch function to speed-up the computation. We initialize the cross-stitch unit to imbalanced, set the standard deviation of the

Gaussian noise to 2, and use simple stochastic gradient descent (SGD) as the optimizer.

All compared methods use the same split as train:dev:test=8:1:1 and the reported results are based on the test set. We use the dev set to tune the threshold for each binary classification problem in the multilabel classification settings of each task.

Experiments ::: Results and Analysis

We report both the micro and macro-F1 scores of the different classification tasks in Tables TABREF27 and TABREF28. Majority refers to labeling based on the majority label, LR to logistic regression, STSL to single task single language models, STML to single task multilingual models, and MTML to multitask multilingual models.

Experiments ::: Results and Analysis ::: STSL

STSL performs the best among all models on the directness classification, and it is also consistent in both micro and macro-F1 scores. This is due to the fact that the directness has only two labels and multilabeling is not allowed in this task. Tasks involving imbalanced data, multiclass and multilabel annotations harm the performance of the directness in multitask settings.

Since macro-F1 is the average of all F1 scores of individual labels, all deep learning models have high macro-F1 scores in English which indicates that they are particularly good at classifying the direct class. STSL is also comparable or better than traditional BOW feature-based classifiers when performed on other tasks in terms of micro-F1 and for most of the macro-F1 scores. This shows the power of the deep learning approach.

Experiments ::: Results and Analysis ::: MTSL

Except for the directness, MTSL usually outperforms STSL or is comparable to it. When we jointly train each task on the three languages, the performance decreases in most cases, other than the target group classification tasks. This may be due to the difference in label distributions across languages. Yet, multilingual training of the target group classification task improves in all languages. Since the target group classification task involves 16 labels, the amount of data annotated for each label is lower than in other tasks. Hence, when aggregating annotated data in different languages, the size of the training data also increases, due to the relative regularity of identification words of different groups in all three languages in comparison to other tasks.

Experiments ::: Results and Analysis ::: MTML

MTML settings do not lead to a big improvement which may be due to the class imbalance, multilabel tasks, and the difference in the nature of the tasks. In order to inspect which tasks hurt or help one another, we trained multilingual models for pairwise tasks such as (group, target), (hostility, annotator's sentiment), (hostility, target), (hostility, group), (annotator's sentiment, target) and (annotator's sentiment, group). We noticed that when trained jointly, the target attribute slightly improves the performance of the tweet's hostility type classification by 0.03,0.05 and 0.01 better than the best reported scores in English, French, and Arabic, respectively. When target groups and attributes are trained jointly, the macro F-score of the target group classification in Arabic improves by 0.25 and when we train the tweet's hostility type within the annotator's sentiment, we improve the macro F-score of Arabic by 0.02. We believe that we can take advantage of the correlations between target attributes and groups along with other tasks, to set logic rules and develop better multilingual and multitask settings.

Conclusion

In this paper, we presented a multilingual hate speech dataset of English, French, and Arabic tweets. We analyzed in details the difficulties related to the collection and annotation of this dataset. We performed multilingual and multitask learning on our corpora and showed that deep learning models perform better than traditional BOW-based models in most of the multilabel classification tasks. Multilingual multitask learning also helped tasks where each label had less annotated data associated with it. Better tuned deep learning settings in our multilingual and multitask models would be expected to outperform the existing state-of-the-art embeddings and algorithms applied to our data. The different annotation labels and comparable corpora would help us perform transfer learning and investigate how multimodal information on the tweets, additional unlabeled data, label transformation, and label information sharing may boost the classification performance in the future.

Acknowledgement

This paper was supported by the Early Career Scheme (ECS, No. 26206717) from Research Grants Council in Hong Kong, and by postgraduate studentships from the Computer Science and Engineering department of the Hong Kong University of Science and Technology.