Twitter Sentiment and Emotion Analysis Technology Corpus

Oihane Cantero
UPV/EHU
ocantero003@ikasle.ehu.eus

Julen Etxaniz
UPV/EHU
jetxaniz007@ikasle.ehu.eus

José Javier Saiz
UPV/EHU
jsaiz014@ikasle.ehu.eus

Abstract

This corpus is a monolingual resource of 140 English tweets about technology. Each tweet is paired with 8 simple adversarial tweets. Both original and adversarial tweets are manually and automatically annotated according to sentiment polarity (positive, negative, neutral) and emotion (joy, anger, optimism, sadness). The objective is to compare and evaluate both types of annotations and measure the impact of adversarial attacks. We created some guidelines to help in manual annotation and saw that there are differences between manual and automatic annotations and also between annotators of the manual annotation.

1 Introduction

This corpus is a monolingual resource of 140 English tweets about technology. Each tweet is paired with 8 simple adversarial tweets, which results in a total of 1120 adversarial examples. Both original and adversarial tweets are manually and automatically annotated according to sentiment polarity (positive, negative, neutral) and emotions (joy, anger, sadness, optimism). For automatic annotation, the TweetEval (Barbieri et al., 2020) evaluation framework was used, which covers both sentiment polarity and emotion detection baseline models for Twitter-specific data, among other classification tasks. The objective is to compare and evaluate both types of annotation.

The aim of this resource is to compare the automatic annotations of tweets about technology and their adversarial sentences. The results will show the performance of this models when tested with new data. They will also show how the model is affected when faced with adversarial samples generated by Adversary[1] library.

[1] https://github.com/airbnb/artificial-adversary

2 Materials and Methods

This section describes the materials and methods that were used to develop the corpus. Data statement, annotation guidelines, Inter Annotator Agreement, automatic annotation and adversarial examples and the resource are described.

The resource and code are publicly available at GitHub[2]. The code is licensed under the MIT open source license. All non-code materials provided are made available under the terms of the CC BY 4.0 license (Creative Commons Attribution 4.0 International license).

2.1 Data Statement

A data statement is a document which serves as the contextualization and description of a data set, usually within the Natural Language Processing context (Bender and Friedman, 2018). It was proposed to make the datasets more understandable for any user and overall, to avoid bad practices within the research and professional practice. In the following sections, we are going to cover the data statement for this corpus to explain the decisions and reasons behind our work.

A. Curation Rationale

The corpus is a monolingual English test set intended to evaluate text classification models for their ability to estimate sentiment and emotion for technology tweets. This is based on their performance against regular documents and adversarial phenomena. It contains a sample of 140 tweets about technology and 8 adversarial examples per tweet. Tweets were obtained from the Twitter API via the Tweepy package. We exclude retweets, quotes and replies to get better tweets. We also exclude tweets that have links to reduce

[2] https://github.com/juletx/twitter-sentiment-emotion

spam tweets. The exact search query used is: *context:65.848920371311001600 lang:en -is:retweet -is:quote -is:reply -has:links*. The goal for using these parameters to collect the text samples was to obtain as many opinions with pronounced sentiments as possible, so that the manual annotation would be as accurate and homogeneous as possible.

B. Language Variety

Data extraction was made using the IETF language tag for English *en* (ISO 639-1). This language label does not refer to any geographic region or specific variety of English, but since the demographic group of speakers is mostly from the United States, as explained in the section 3.1-C, the main language code is *en-US* (ISO 639-1/ ISO 3166-1 alpha-2), which refers to all American English varieties.

C. Speaker Demographic

Tweets are included upon language and semantic content. Any other type of information or demographic data cannot be controlled for selection, which means that this corpus is not based on specific speakers, but on randomly chosen instances from a larger collection. Therefore, we must rely on general Twitter demographics, which include a majority of speakers which are male (70.4%) and ranges between 18 and 49 years old (78.7%). Tweet metadata sometimes provides location information, which for our sample suggests that a significant number of the instances originate in the United States. This information is relevant for determining the prevailing language variety in the corpus.

D. Annotator Demographic

The annotation was made by the authors of this paper, three Spanish MSc students with training in linguistics and computer science.

E. Speech Situation

Tweets are characterized by being short, spontaneous and conversational in nature. That is, opinions are concise and sentiments are condensed into a few words, which is important for the corpus because it guarantees that the annotation will be carried out and analyzed accurately. Tweets are also time-sensitive texts, which means that the content the instances is relevant to the time in which they were written.

F. Text Characteristics

The sample includes tweets made before February 1 2022. The Twitter context *Interests and Hobbies Vertical: Technology* (domain and entity ID: 65.848920371311001600) was used as the thematic parameter. This context includes ”top-level interests and hobbies groupings” about ”technology and computing”, as indicated in the description of the metadata context. Check Figure 1 to have an idea of the most frequent words in the tweets.

Figure 1: Wordcloud of all tweets.

2.2 Guidelines

We defined very simple initial guidelines before starting the annotation process. The tweets were annotated according to the sentiment (positive, negative, neutral) and emotion (joy, anger, sadness, optimism) that was inferred by the semantic content. Only one label per tweet was allowed for both tasks. We read tweets one by one and annotate both tags.

After the first annotation, we realized that we needed to update the guidelines to increase the annotator agreement. These guidelines are summarized and exemplified in Table 1.

On the one hand, there were some misunderstandings with words and symbols. For example, <3 is used as a heart and > equals greater than. On the other hand, there were some were some cases where specific criteria had to be decided to select a label.

When annotating sentiments, some tweets seemed to be ironic or sarcastic, and it was difficult to decide a tag. Since irony makes the ”intended meaning [...] appear on the surface to express the opposite” (Irony, 2022), we decided for those cases

Problem	Guideline	Example	
		Sentence	Label
Irony/ Sarcasm	Use opposite sentiment to the literal one	”oh yeah tesla well what about a car that just logs into your tiktok acct and drives you to starbucks”	negative
Contradicting sentiments	Prefer overall sentiment over the ”neutral” tag	”Finally managed to move my business email from google hosting to another host. So stressful and difficult. The whole internet is so hard! At least the metaverse is coming, I'm confident that will make everything easy and good, phew.”	positive
Unclear emotion	Prefer ”joy” tag for being broader in sense	”All the software I create will be free and open source, but that doesn't necessarily mean I won't write cryptic software for some of my projects”	joy

Table 1: Defined guidelines after the Inter Annotator Agreement with corresponding examples.

expressing irony or sarcasm to label the opposite sentiment to the overall sentiment when taken literally.

Another problem for which a clear guideline was needed was tweets with contradicting sentiments, tweets that have both positive and negative sentiments. For this case, we decided to keep the overall sentiment, assuming that there is always one sentiment that is stronger than the other, to avoid using too many neutral tags.

When there was no clear emotion, we decided to select the *joy* tag because it is the term with the broadest meaning among the possible options. Also, we found that we had some uncertainty between the *joy* and *optimism* emotions, as they are the two ”positive” emotions we had, and sometimes we weren't sure which one was more appropriate, so we also decided to select *joy* for the same reason.

2.3 Inter Annotator Agreement

Initially, a sample of the first 20 instances was shared among the annotators. The annotations for these initial tweets were used to calculate Inter Annotator Agreement. Table 2 shows examples of these instances.

Table 3 shows that there is moderate agreement and a low Kappa score in the first results. The agreement results in the sentiment task are homogeneous between pairs, while in the emotion task the results were more varied. This is explained by the task labels: the sentiment task uses labels which are more clear, while the labels in the emotion task are more subjective and open to interpretation. This is the main reason why the guidelines were updated, after which the annotation was repeated along the Inter Annotator Agreement calculation. The results for the annotations with the updated guidelines show a clear improvement in the emotion task.

Thereafter each annotator was given a different set of 40 tweets for annotation. The same methodology was followed, where the manual annotations were used to evaluate the accuracy of the models.

2.4 Automatic Annotation

All the tweets were also labelled automatically using the baseline models trained for the TweetEval (Barbieri et al., 2020) evaluation framework. This framework covers many classification tasks that include sentiment and emotion classification.

The base RoBERTa (Liu et al., 2019) model was pretrained using 58M English tweets. Then each model is finetuned using specific data for that task. SemEval 2017 Subtask A data (Rosenthal et al., 2019) is used for sentiments and SemEval 2018 Affects in Tweets (Mohammad et al., 2018) is used for emotions.

The models are available in HuggingFace Transformers library and are very easy to use. They only require a small preprocessing step to substitute all mentions with @user and links with http.

Text	Sentiment			Emotion		
	Julen	**Oihane**	**Javier**	**Julen**	**Oihane**	**Javier**
just put a CD into my MacBook to burn it and my computer is literally trembling with reawakened recognition	negative	positive	negative	anger	joy	joy
oh yeah tesla well what about a car that just logs into your tiktok acct and drives you to starbucks	neutral	positive	negative	joy	joy	anger
#100DaysOfCode Haven't updated in a while due to not feeling well, just been reviewing some HTML/CSS & JavaScript until I feel better to take on new concepts. Also been watching mock interviews :)	negative	negative	negative	sadness	optimism	optimism

Table 2: Instance examples used to calculate Inter Annotator Agreement.

pairs	agreement_sentiment	kappa_sentiment	agreement_emotion	kappa_emotion
Julen Oihane	70.0 → **65.0**	53.8 → **45.1**	55.0 → **70.0**	33.3 → **49.2**
Julen Javier	60.0 → **70.0**	40.1 → **49.4**	40.0 → **70.0**	13.0 → **46.9**
Javier Oihane	60.0 → **55.0**	39.2 → **32.3**	65.0 → **70.0**	46.2 → **50.0**
average	63.3 → **63.3**	44.4 → **42.3**	53.3 → **70.0**	30.8 → **48.7**

Table 3: Inter Annotator Agreement result improvements after updating the guidelines. Final results are in bold.

2.5 Adversarial Examples

We created some adversarial tweets automatically to measure the impact in the metrics of the models. The aim of these adversarial examples is to confuse the models while maintaining the manual label. Therefore, if we generate the adversarial examples correctly, they should have the same meaning and sense, so there is no need to annotate them again manually.

We used the Adversary[3] library to generate 8 generic attacks that include swapping words or letters among others. These attacks are not specifically prepared for our models, they are simple attacks that can be tested on any type of model. Words are selected with a probability of 0.3 and the selected attack is applied to those words. Attacks can also be combined, but we decide to apply them separately for simplicity.

An example of each attack can be seen in Table 4, and we can compare the manual annotation with all the adversarial annotations. We can first see that for the original tweet, the manual and automatic annotations doesn't match for emotion. And some adversarial attacks change the sentiment from negative to neutral and the emotion from sadness to anger.

[3]https://github.com/airbnb/artificial-adversary

2.6 Resource

The resource is available at GitHub and it is composed of six files. We decided to separate the resource into these six files to avoid having a big file with all the information that would be unintelligible and with a large number of columns. ITA annotation are also kept in two separate files that correspond to different annotations.

- *tweets.csv* : The tweets with the manual and automatic annotations for sentiment analysis and emotion detection.

- *tweets_ita.csv*: The first 20 tweets with the annotations made by each of the annotators for both tasks with the initial guidelines. It is used

Type	Sentence	Label	
		Sentiment	**Emotion**
Manual Original	having no sort of WiFi actually sucks	negative	anger
Automatic Original	having no sort of WiFi actually sucks	negative	sadness
Swap words	having no sort WiFi of actually sucks	negative	sadness
Remove space	having-no‘sort of WiFi actually sucks	negative	anger
Replace letters with symbols	having no $or7 of WiFi actually $u{[}kS	neutral	sadness
Swap letters	having no srot of WiFi actually sucks	negative	anger
Insert punctuation	having no sor]{t of WiFi actually]su!cks	neutral	sadness
Insert duplicate characters	having no sorttt of WiFi actually sucks	negative	sadness
Delete characters	having no sot of WiFi actually sucks	negative	sadness
Change case	having no SORT of WiFI actually sucks	negative	anger

Table 4: Examples of adversarial sentences and the predicted sentiment and emotion.

to calculate the Inter Annotator Agreement with the first guidelines.

- *tweets_ita2.csv*: The first 20 tweets with the annotations made by each of the annotators for both tasks with the updated guidelines. It is used to calculate the Inter Annotator Agreement once we updated the guidelines.

- *tweets_adv.csv*: The original tweets and the adversarial tweets.

- *tweets_adv_sent.csv*: The result of the manual annotation and the automatic annotation for the original tweets and all the adversarial tweets for sentiment analysis.

- *tweets_adv_emot.csv*:The result of the manual annotation and the automatic annotation for the original tweets and all the adversarial tweets for emotion detection.

3 Results

We compare the F1 scores and extract some conclusions about the performance of the models. The F1 scores of each task, tag and adversarial attack are compared. We also visualize the percentage of tweets from each class in the manual and automatic annotations. We try to explain the main differences between manual and automatic annotation.

3.1 F1 Scores

F1 scores were calculated for each label and macro-average for all labels in each task in order to compare the model results against the manual annotations. As can be seen in Table 5, the results with adversarial sentences output lower F-scores than with the original sentences. For the sentiment task, the performance with original sentences delivers a 65.8 F-score against a 61.7 mean F-score with the adversarial sentences. The same applies to the emotion task results, which bring a 53.3 F-score with original sentences and a 50.7 mean F-score with the adversarial sentences.

We also see that in sentiment analysis, the results for the three tags are nearly equivalent (around 60-70) for every type of sentences. In emotion detection, the scores for anger and joy are also quite similar (also around 60-70), the ones for sadness are a little worse (around 40), but the ones for optimism are very bad for both original and adversarial sentences (around 25).

Even though overall results for adversarial sentences are worse than the original ones, we can remark that some of the adversarial attacks doesn't decrease F-score, they even improve it in some cases. The best results for sentiment analysis are for the *swap_words* attack and this same attack performs quite well in emotion detection. And for emotion detection the best results are for *remove_spacing*, that doesn't receive so bad results for sentiment analysis. So watching these results, we may conclude that removing a space of the sentence or word order doesn't really matter for these tasks.

In Figure 3, we see the results of Table 5 as barplots, we see the score differences between each tag. The tag that is classified worse for both original and adversarial sentences is *positive*, and for emotions, we see that *anger* and *joy* are the tags that get better results, and *optimism* is clearly below all the others.

In Figure 4 we can observe the results for each task and attack. Sentiment analysis gets better re-

text	neg	neu	pos	sen	ang	joy	opt	sad	emo
original	71.0	71.0	70.3	65.8	66.7	66.1	27.9	52.6	53.3
swap_words	**73.1**	**73.1**	**71.0**	**67.2**	67.6	64.6	27.9	51.3	52.8
remove_spacing	68.2	68.2	60.2	61.2	**69.6**	**68.2**	33.3	**55.0**	**56.5**
letter_to_symbol	65.2	65.2	58.5	59.0	62.5	62.0	22.9	38.9	46.6
swap_letters	69.0	69.0	64.3	62.8	67.6	64.1	21.6	31.6	46.2
insert_punctuation	62.8	62.8	55.7	57.5	62.5	74.0	20.0	35.0	47.9
insert_duplicate_characters	72.5	72.5	63.6	63.9	66.7	66.7	26.3	47.4	51.8
delete_characters	68.2	68.2	59.5	59.9	68.5	66.7	27.8	41.0	51.0
change_case	71.1	71.1	63.6	62.6	63.2	65.6	**37.2**	44.4	52.6

Table 5: F1 scores for original text and adversarial attacks. F1 for each label and macro-average of all labels for sentiment and emotion.

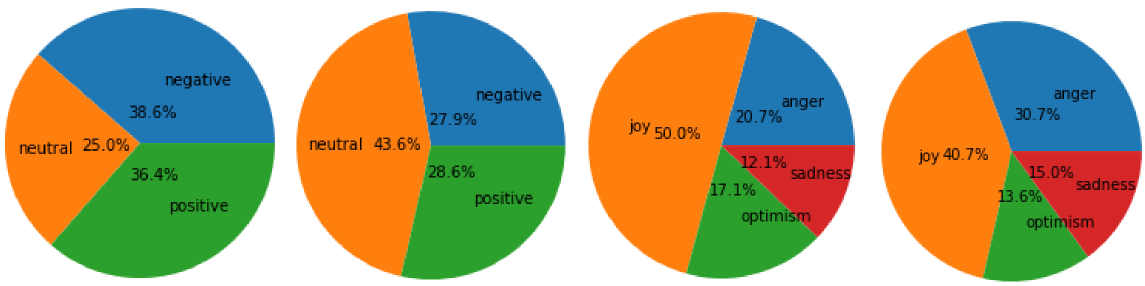

Figure 2: Sentiment and emotion label percentages for manual and automatic annotation.

sults for both original and adversarial sentences. Figure 4 gives an idea of the distribution of the scores for each task and attack. Sine attacks get bad results in both tasks while others get good results in both tasks. There are also some attacks that have much worse scores for one task than the other.

3.2 Pieplots

We drew some pieplots that can be seen in Figure 2 to compare the distribution of the tags of each of the tasks. There are some big differences that are worth noting about the tag distributions.

The two first plots correspond to the distribution of each of the tags of the sentiment analysis task for manual and automatic annotation. In the manual distribution, positive and negative tags are equally distributed, and are less neutral tags. In the automatic annotation, the distribution of positive and negative tags are also equally distributed, but the majority class is neutral. This can be because while manually annotating, we tend to avoid neutrality, and we try to put a positive or negative tag at all costs. Another reason may be that we decided to take the overall sentiment when there are positive and negative sentiments in the same tweets, and maybe the model makes an average and classifies them as neutral.

The last two plots correspond to emotion tag percentages for manual and automatic annotation. In our guidelines, we decided that we will put the joy tag when we will have doubts. This increased our manual annotations' joy tags number. For optimism and anger, the results are quite similar but for anger, because the automatic annotation puts more than the manual one. This should be because our guidelines implies that there will be more joy, because we put it when there is no emotion and when we had doubts between joy and optimism, as we think that it is the more general emotion.

4 Conclusions

Classification tasks in Twitter related to sentiment analysis have grown in importance over the last years, and with an increasing number of models and frameworks created around this tasks, it is worth evaluating the implementation of these methods.

Our evaluation focuses on the outcomes of a popular Twitter-specific framework across two tasks by comparing its results to human annotation. This evaluation method is useful because it allows us

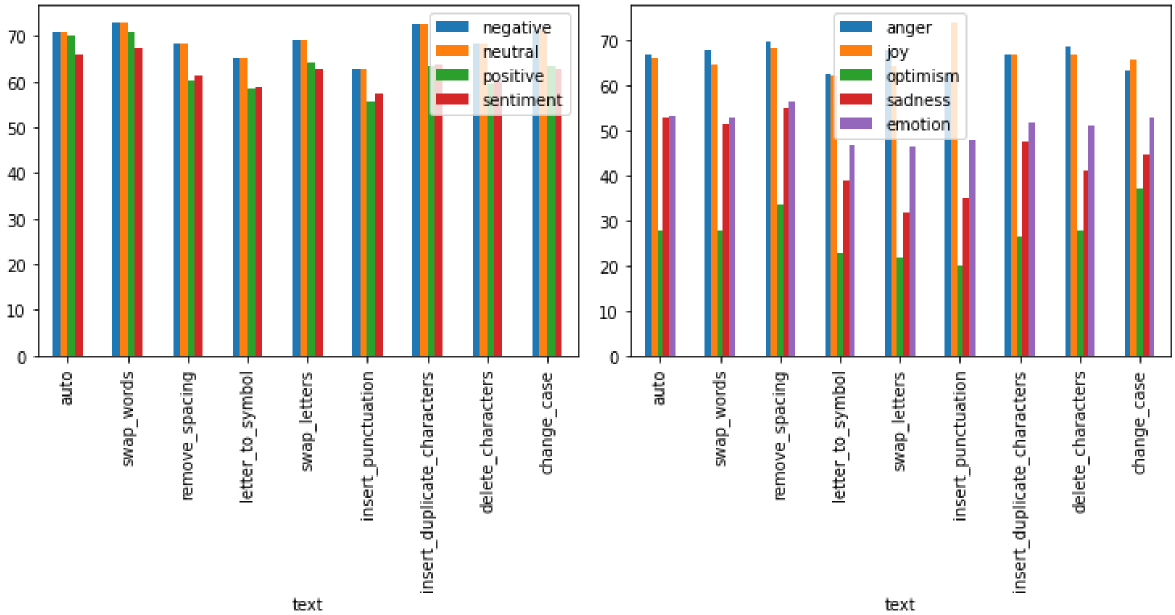

Figure 3: F1 for each label and macro-average of all labels for sentiment and emotion.

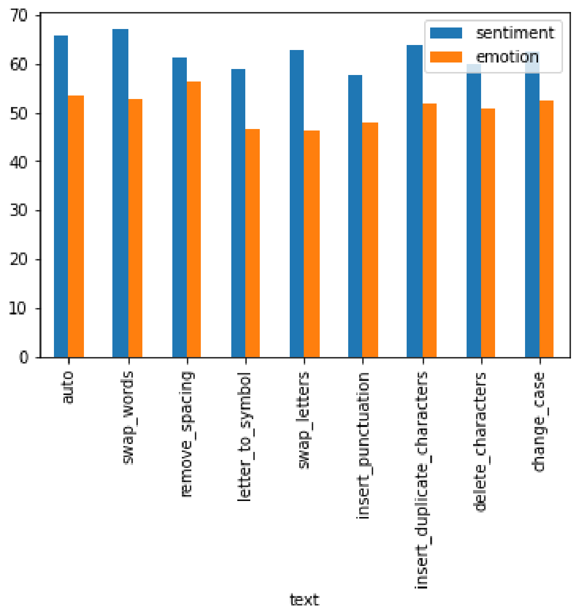

Figure 4: Sentiment and emotion barplot.

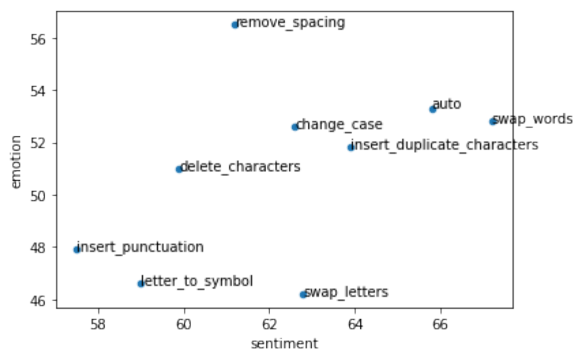

Figure 5: Sentiment and emotion scatterplot.

to examine the performance of these models when tested with different data.

Testing models against adversarial examples is necessary to find the weaknesses of these models. We saw that very simple adversarial examples can alter the performance of large models. Many of these attacks are unnoticeable for humans, as they can be seen as typical errors made on Twitter.

We saw that good guidelines are essential to annotate manually, even if the task seems to be ”easy”. They are very important if we want to create a coherent resource. Annotation errors have a big impact for training and testing these models and lead to incorrect results.

The outcomes of this work have shown that classification models have multiple sources of error and tend to under-perform when facing different scenarios such as unknown words, lexicon gaps or unspecified sentence contexts. This shows there is room for improvement.

For future work, it would be useful to analyze the performance of the models in multiple other related tasks, such as irony/sarcasm detection or hate detection, instead of a pair of tasks. Other improvements on this evaluation would be to explore additional languages and other classification frameworks with different guidelines.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Irony. 2022. *Oxford Reference*. Oxford University Press.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17, New Orleans, Louisiana. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2019. Semeval-2017 task 4: Sentiment analysis in twitter. *arXiv preprint arXiv:1912.00741*.