Twitter Sentiment and Emotion Analysis

Oihane Cantero
UPV/EHU
email@domain

Julen Etxaniz
UPV/EHU
email@domain

Jose Javier Saiz
UPV/EHU
email@domain

Abstract

This corpus is a monolingual resource of unique tweets about product reviews, with each tweet paired with an adversarial sentence. Both original and adversarial tweets are manually and automatically annotated according to sentiment polarity (positive, negative, neutral). The objective is to compare and evaluate both types of annotation.

1 Introduction

This corpus is a monolingual resource of unique English tweets about technology, with each tweet paired with an adversarial sentence. 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 stance detection 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 annotations of tweets about technology and their adversarial sentences. It will also be interesting to compare these annotations when we do it by hand or automatically. The results will show whether the models are affected in their performance on classification tasks when faced with adversarial samples. If performance is significantly compromised by this phenomenon, improvements in the criteria used to generate the models may be proposed.

2 Related work

3 Material and methods

3.1 Data statement

(Bender and Friedman, 2018)

The resource and code are publicly available at GitHub[1]. 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).

A. CURATION RATIONALE

The corpus is a monolingual test set intended to evaluate text classification models for their ability to estimate sentiment and emotion. This is based on their performance against regular documents and adversarial phenomena. It contains a sample of 140 tweets about technology 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 spam tweets. The exact query used is: *context:65.848920371311001600 lang:en -is:retweet -is:quote -is:reply -has:links*

B. LANGUAGE VARIETY

Data extraction was made using the BCP 47 identifier for English (ISO 639-1 en). Since this tag does not address any specific variety of English and Twitter is used worldwide, we can assume that the extracted sample includes a number of varieties of English.

C. SPEAKER DEMOGRAPHIC

Because we can not infer any precise information about the authors from the text alone, 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 geolocation information, which for our sample suggests that a significant number of them originate in the United

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

States. Beyond that, our sample demographics are only restricted by language and topic of interest.

D. ANNOTATOR DEMOGRAPHIC

The annotation was made by three Spanish MSc students with training in linguistics and informatics.

E. SPEECH SITUATION

Tweets are characterised by being short, spontaneous, time-sensitive texts. This means that the content of tweets is only relevant to the time in which they were written.

F. TEXT CHARACTERISTICS

The sample includes tweets made before February 1. 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.

3.2 Guidelines

Only one label per tweet is allowed. We will annotate the tweets depending on their sentiment (positive, negative, neutral) and emotion (joy, anger, sadness, optimism).

3.3 Inter Annotator Agreement

Initially, a sample of the first 20 instances will be shared among the annotators. These initial tweets will be used to calculate Inter Annotator Agreement and make modifications to the guidelines and scope if necessary.

We will calculate the ITA between us three and the automatic results for the 20 shared tweets. We might also calculate the agreement between each of us and the automatic results for the 40 tweets we will annotate alone.

We will compute the Kappa scores to see if we agree in the annotations of both sentiment and emotions, and if we agree with the automatic annotations.

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

3.4 Automatic annotation

These tweets will also be labelled using automatic models. Two HuggingFace models that are fine-tuned for each task will be used for evaluation. We will compare the results with our annotations to have an idea of the accuracy of our models.

Finally, we will compare the results and extract some conclusions. We will visualize the percentage of tweets from each class in the manual and automatic annotations. We will also visualize the most common words of each class in word clouds.

3.5 Adversarial examples

We will create 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, there is no need to annotate them. As these attacks can easily be replicated, we decided not to include them in the final resource.

We used Adversary[2] library to generate 8 generic attacks. These attacks are not specifically prepared for our models, they are simple attacks that can be tested on any type of model.

1. Swapping words (swap_words)
2. Removing spacing between words (remove_spacing)
3. Replacing letters with similar-looking symbols (letter_to_symbol)
4. Swapping letters (swap_letters)
5. Inserting punctuation (insert_punctuation)
6. Inserting duplicate characters (insert_duplicate_characters)
7. Deleting characters (delete_characters)
8. Changing case (change_case)

An example of each attack can be seen in Table...

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

4 Results

4.1 Inter Annotator Agreement

4.2 Automatic annotation metrics

5 Conclusions

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.