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

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 data sets 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 so to explain the decisions and reasons behind our work.

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 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.

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

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 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 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 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

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 is 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. 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 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. For this case, we decided to keep the overall sentiment to avoid using too many neutral tags.

When there is no clear emotion, we decided to select the *joy* tag because it is the term with the broadest meaning among the possible options. Also, if there is uncertainty between the *joy* and *optimism* sentiments, we also decided to select *joy* for the same reason.

These guidelines are summarized and exemplified in Table 1.

3.3 Inter Annotator Agreement

Initially, a sample of the first 20 instances was shared among the annotators. These initial tweets were used to calculate Inter Annotator Agreement, after which we made modifications to the guidelines and scope of the work.

The ITA between the three annotators and the automatic results was calculated for the 20 shared tweets. We also calculated the agreement between each annotator and the automatic results for the 40 tweets that were annotated individually.

The Kappa scores were computed to see if we agreed in the annotations of both sentiment and

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.

emotions, and if we agreed with the automatic annotations.

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.

3.4 Automatic annotation

These tweets were also labelled using automatic models. Models trained for the TweetEval (Barbieri et al., 2020) evaluation framework were used. 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 fine-tuned models are available in HuggingFace and that makes them very easy to use. They only require a small preprocessing step to substitute all mentions with @user and links with http.

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, they should have the same meaning and sense, so 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 the 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. 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 2. This tweet was manually annotated as negative sentiment and anger emotion.

3.6 Visualization

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.

4 Results

First, we will analyse ITA results in the first and second annotation. Then we will compare the results and extract some conclusions about the performance of the models.

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

text
just put a CD into my MacBook to burn it and my computer is literally trembling with reawakened recognition
oh yeah tesla well what about a car that just logs into your tiktok acct and drives you to starbucks
#100DaysOfCode Haven't updated in a while due to not feeling well, just been reviewing some HTML/CSS & Ja

Type	Sentence	Label	
		Emotion	Sentiment
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 2: Examples of adversarial sentences and the predicted sentiment and emotion.

4.1 Inter Annotator Agreement

The first results of Inter Annotator Agreements were quite low, so we realized that we needed new guidelines. After updating the guidelines, we got better results.

4.2 Automatic annotation metrics

5 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 for the purpose of this work because it allows us to examine the broader potential impact of these models and whether the guidelines used to create the training corpus for these tasks translate into adequate performance on sentiment analysis tasks.

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 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*.