Submission for
Exercise task 2
of UTU course TKO_8964-3006
Textual Data Analysis
by Botond Ortutay

Instructions:

The Stanford Sentiment Treebank (SST) dataset which we used as training data in the explainability part of the lecture has more to it than it seems at first sight, and is a good example of a dataset constructed with some thought and deep understanding of the underlying task. It was introduced in this 2013 paper: https://aclanthology.org/D13-1170.pdf

Read the parts of the paper which are relevant to the construction of the SST dataset and the evaluation (Sections 3&5). The paper also introduces some models which of course from today's point of view seem a mere historical curiosity, but you can check those out too.

Ponder/answer the following:

1. What is the distinguishing feature of the SST corpus, as opposed to other sentiment-labeled corpora?
2. Who were the annotators?
3. Does the paper give enough information to establish whether the annotators were in good agreement and whether you can trust the dataset?
4. Google the authors and check their Google Scholar, several of them are true stars of the field!

Answers:

What is the distinguishing feature of the SST corpus, as opposed to other sentiment-labeled corpora?

The most important thing seems to be the storing of sentences (or phrases) in trees, where each word is a leaf in the tree structure, and each leaf has a sentiment value by itself, and the final sentiment value of the whole tree (sentence) is constructed from the sentiments of the leaves (words) and branches (phrases).

This allows the more in depth grammatical analyses they were performing in the article (contrastive conjunction and high level negation analyses).

Who were the annotators?

To quote the article:

> It was parsed with the Stanford parser (Klein and Manning, 2003) and includes a total of 215,154 unique phrases from those parse trees, each annotated by 3 human judges.

In other words: the article doesn't exactly name drop the annotators, it only says that each parse tree was annotated by several humans (that is my interpretation of the above text anyway). If I had to guess it was either somebody involved with the project (possibly interns or something) or it was outsourced in some way.

Does the paper give enough information to establish whether the annotators were in good agreement and whether you can trust the dataset?

The article didn't really talk about the annotators that much, mostly focusing on the tools the annotators used and the data the annotators gave back. As such I can't entirely be sure that they didn't just pay some company to crowdsource the results with tools given to them. So it's not entirely sure that the annotators weren't just randoms from the street who did the annotations for fun or for 5$. But still, as this is a large amount of data, and I still assume that most of the annotations are "of high quality" so to speak and given that there were multiple annotations for each phrase, I'd say that I'd trust the database overall.

I can't really compare on whether there were many disagreements on sentiments of phrases or not. I mean how negative a phrase seems is personal, and even if all the annotators were 100% focused on the task and were 100% honest they'd still probably give different results.

For example: Table 3 (page 10) contains the most negative phrases a model trained on the dataset (probably the train split) found in the dev split. One of these phrases was "silliest and most incoherent movie" which I wouldn't consider that negative. Certainly not in the top 5 most negative 5-word phrases found in a 1101-sentence dataset; All that is to say there probably were minor disagreements, and that's probably why they got several opinions for the sentimentality of each phrase anyway. There was probably some mathematical model calculating weighted scores per phrase in the background. Maybe with the sentences with several contradictory sentiments somebody on the project gave a better look?

Google the authors and check their Google Scholar, several of them are true stars of the field!

- Richard Socher certainly has a lot of articles. Apparently he was also involved in the creation of Imagenet, which I've heard of.
- Alex Perelygin on the other hand was only involved in 4 articles and only this one is noteworthy at all. He was probably a student at Stanford when he was involved in this project.
- Jean Y Wu similarly has just a few publications (16), and this is also the only major article she's been involved in.
- Jason Chuang has a few articles on visual analyzation & such but this is his most cited publication.
- Christopher D. Manning is an established researcher (in fact I think I've heard of him already), with several publications cited over 10000 times. He also seems to have written a book on information retrieval for Cambridge University.

- Andrew Y. Ng also has several publications cited over 10000 times. He seems to have been involved in the creation of ROS (Robot Operating System), another project I'm familiar with.
- Christopher Potts is a professor of linguistics in Stanford University. He has truly a lot of publications (100+) but this article is his most influential. If I had to guess he was probably involved in the grammatical analysis part of this project.

On sources and generative AI use:

All sources used have been mentioned in the instructions (linked article + Google Scholar pages of authors). As such I don't feel the need to include a proper references section.

No generative AI tool has been used in any of the answers above in any way.