Information Retrieval and Web Search Engines
Contextual Model for Sentiment Extraction from Tweets

Toygar Tanyel
18011094

28.12.2021

Abstract

Sentiment Analysis has an important role in today's world especially for private companies which hold lots of data. The massive amount of data generated by Twitter present a unique opportunity for sentiment analysis. However, it is challenging to build an accurate predictive model to identify sentiments, which may lack sufficient context due to the length limit. In addition, sentimental and regular ones can be hard to separate because of word ambiguity. In this project, I will be proposing the phases of text pre-processing, visual analysis, modeling, and Twitter scraping.

1 Introduction

Social media has been increasingly popular for people to share instant feelings, emotions, opinions, stories, and so on. As a leading social platform, Twitter has gained tremendous popularity since its inception. The latest statistical data show that as of May 2020, on average, around 6,000 tweets are sent in every second. Society generate a massive amount of social data which are used by numerous upper-level analytical applications to create additional value. Therefore, enough data is generated for sentiment analysis to give an intuition us to understand intention of people. The tweets are not clean. That means we will start with the data cleaning, and continue with the changing dataset in order to make it useful for our purpose. BERT, CNN and Bidirectional LSTM will be used and explained later in the paper.

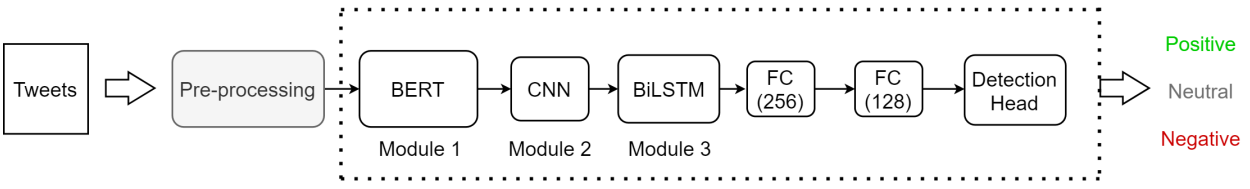

Figure 1: The proposed BERT-CNN-BiLSTM-FC learning pipeline for sentiment prediction. FC (Fully Connected Layer)

The rest of this paper is organized as follows: Section 2 covers the dataset description and shows the statistics of data; Section 3 explain the technical details of the proposed learning model; Section 4 includes details about information retrival process; Section 5 provides experimental validation with information retrieval and result analysis; Section 6 summarizes my work.

2 Material and Methods

Entire code in notebook for this task can be found here. The sentiment140 dataset is used. It contains 1,600,000 tweets extracted using the twitter API. The tweets have been annotated (0 = negative, 4 = positive) and they can be used to detect sentiment. The dataset was created by [2].
The way of the collection of dataset according to the creators:
"Our approach was unique because our training data was automatically created, as opposed to having humans manual annotate tweets. In our approach, we assume that any tweet with positive emoticons, like :), were positive, and tweets with negative emoticons, like :(, were negative. We used the Twitter Search API to collect these tweets by using keyword search".

The data is a CSV with emoticons removed. Data file format has 6 fields:

- target: the polarity of the tweet (0 = negative, 4 = positive), however, we will change it as (0 = negative, 2 = neutral, 4 = positive) to create one more alternative sentiment
- ids: The id of the tweet
- date: the date of the tweet
- flag: The query (lyx). If there is no query, then this value is NO_QUERY.
- user: the user that tweeted
- text: the text of the tweet

Accessed data from Kaggle—Sentiment140 dataset with 1.6 million tweets

	target	ids	date	flag	user	text
0	0	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...
1	0	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...
2	0	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...
3	0	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire
4	0	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....

Figure 2: Raw Dataset

2.1 Data Pre-Processing

Dataset includes 2 specific classes (0 - negative, and 4 - positive). To be able to make "Neutral" class usable as an alternative with threshold, dataset will be mapped manually. More precisely, the mapping means is that creating a dictionary for labels as ({0: NEGATIVE, 2: NEUTRAL, 4: POSITIVE }). Afterwards, the urls, html tags and punctuations will be removed to decrease complexity of BERT's own pre-processing step during tokenize the texts.

	target	ids	date	flag	user	text	text_clean
0	NEGATIVE	1467810369	Mon Apr 06 22:19:45 PDT 2009	NO_QUERY	_TheSpecialOne_	@switchfoot http://twitpic.com/2y1zl - Awww, t...	switchfoot Awww thats a bummer You shoulda ...
1	NEGATIVE	1467810672	Mon Apr 06 22:19:49 PDT 2009	NO_QUERY	scotthamilton	is upset that he can't update his Facebook by ...	is upset that he cant update his Facebook by t...
2	NEGATIVE	1467810917	Mon Apr 06 22:19:53 PDT 2009	NO_QUERY	mattycus	@Kenichan I dived many times for the ball. Man...	Kenichan I dived many times for the ball Manag...
3	NEGATIVE	1467811184	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	ElleCTF	my whole body feels itchy and like its on fire	my whole body feels itchy and like its on fire
4	NEGATIVE	1467811193	Mon Apr 06 22:19:57 PDT 2009	NO_QUERY	Karoli	@nationwideclass no, it's not behaving at all....	nationwideclass no its not behaving at all im ...

Figure 3: New dataset, mapped and cleaned

2.2 Visualizing the Data

Visualizing the data benefits us in many ways. Visuals make analysis easier and faster, while giving you the ability to see important topics at a glance. Therefore, understanding and thinking the next step become more clear. We have relatively big data which includes 1.6 millions of tweets that makes visualization quite crucial.

Dataset has distribution as follows:

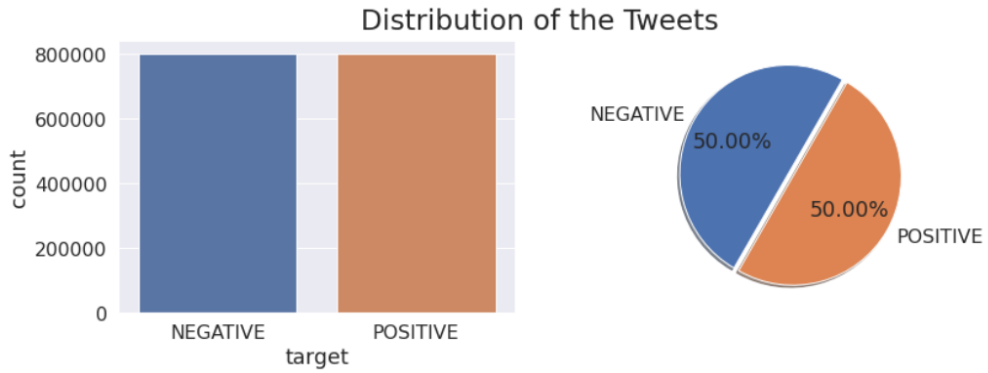

Figure 4: Distribution of the Dataset

As a quick and easy observation, we can say that dataset has no imbalanced label problem. Negative and Positive labels are equally distributed. The situation of equilibrium will let model to learn more accurate. However, should not be forgotten, dataset might has lots of mislabelled text due to way of collection which has only parameter as *":)" : positive* or *":(" : negative*. The problem is that lots of people use those kind of representations ironically.

Character count is used often as a part of visual analysis for natural language processing. The number of characters might give an intuition about the difference between texts.

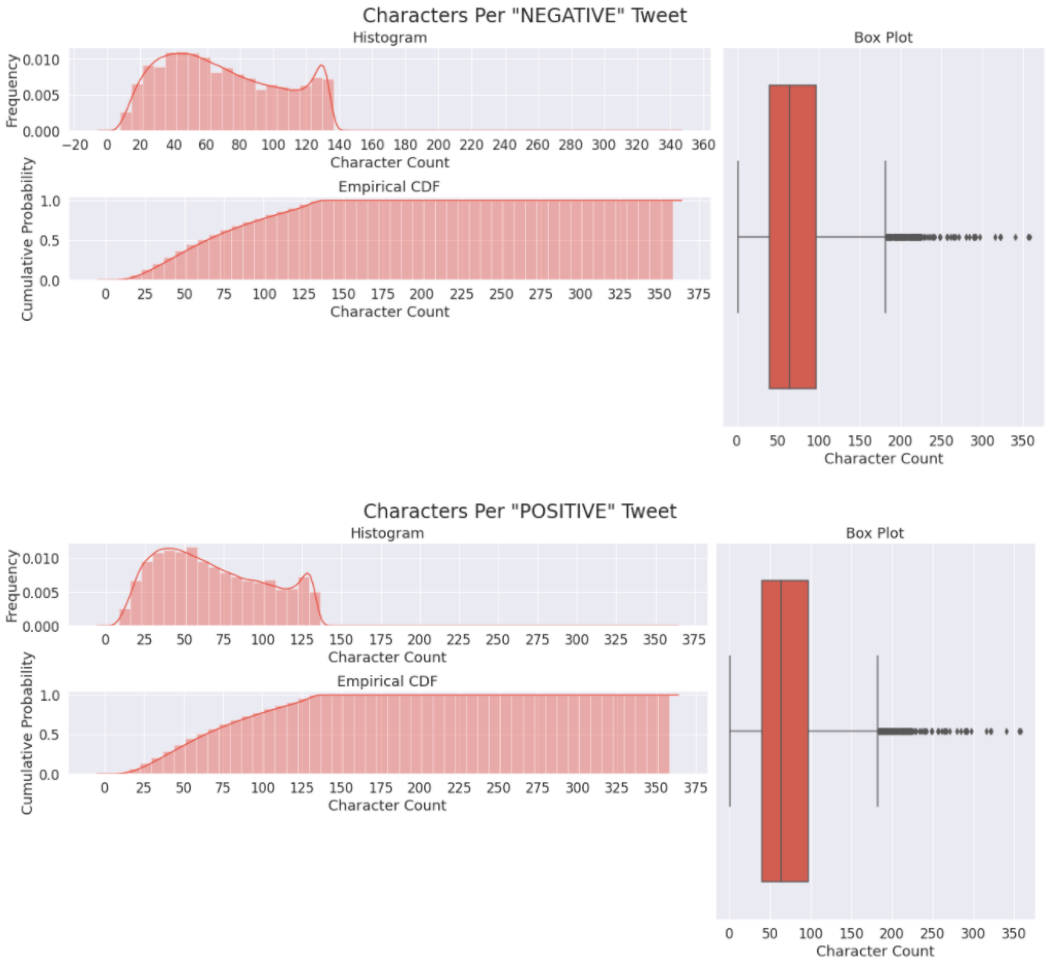

Figure 5: Character count of negative and positive labelled texts.

3 Model

Figure 1 shows the proposed BERT-CNN-BiLSTM-FC learning pipeline, which consists of four sequential module. BERT [1] produces competitive results, and can be considered as the new electricity of natural language processing tasks such as sentiment analysis, named entity recognition (NER), and topic modeling. The combination of CNN and BiLSTM models requires a particular design, since each model has a specific architecture and its own strengths:

- BERT is utilized to transform word tokens from the raw Tweet messages to contextual word embeddings.
- CNN is known for its ability to extract as many features as possible from the text.
- BiLSTM keeps the chronological order between words in a document, thus it has the ability to ignore unnecessary words using the delete gate.
- Fully Connected Layers give robustness to decrease unsteadiness of results in hard cases.

The purpose of combining CNN and BiLSTM models is to create a model that takes advantage of the strengths of both, so that it captures the features extracted using CNN, and uses them as an BiLSTM input. Therefore, I develop a model that meets this objective, such that the vectors built in the word embedding part are used as convolutional neural network input. Then takes features as input of BiLSTM layers. Afterwards, FC takes the output of BiLSTM to make weights more robust. The output of the FC is fed to a detection layer to generate the final prediction result, i.e., positive, negative (or neutral).

3.1 BERT (Bidirectional Encoder Representations from Transformers)

Google presented a novel neural network architecture called a transformer in *Attention is all you need* [5] which had many benefits over the conventional sequential models (LSTM, RNN, GRU etc). Advantages included but were not limited to, the more effective modeling of long term dependencies among tokens in a temporal sequence, and the more efficient training of the model in general by eliminating the sequential dependency on previous tokens. BERT is an attention-based language model that utilizes a stack of Transformer encoders and decoders to learn textual information. It also uses a multi-headed attention mechanism to extract useful features for the task. The bidirectional Transformer neural network, as the encoder of BERT, converts each word token into a numeric vector to form a word embedding, so that words that are semantically related would be translated to embeddings that are numerically close.

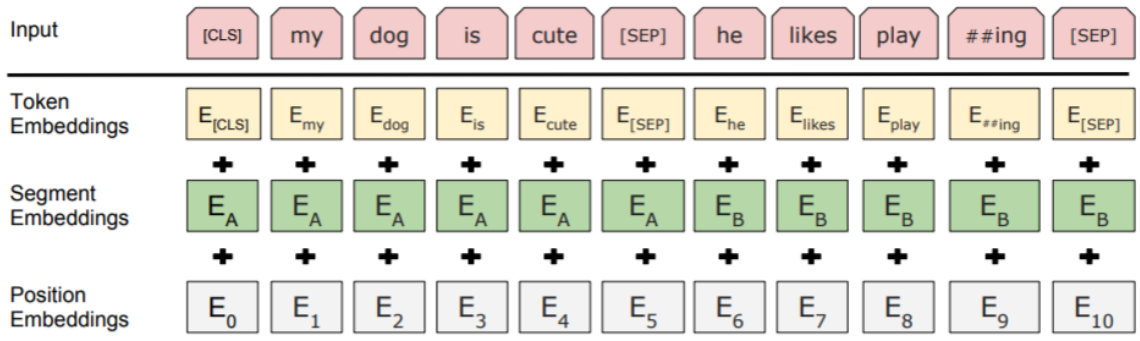

Figure 6: BERT input representation. The input embeddings are the sum of the token embeddings, the segmentation embeddings and the position embeddings.

3.2 CNN (Convolutional Neural Networks)

CNN is a class of deep, feed-forward artificial neural networks (where connections between nodes do not form a cycle) & use a variation of multilayer perceptrons designed to require minimal preprocessing. [4] shows that a simple CNN with little hyperparameter tuning and static vectors achieves remarkable results on multiple benchmarks. Learning task-specific vectors through fine-tuning offers further gains in performance.

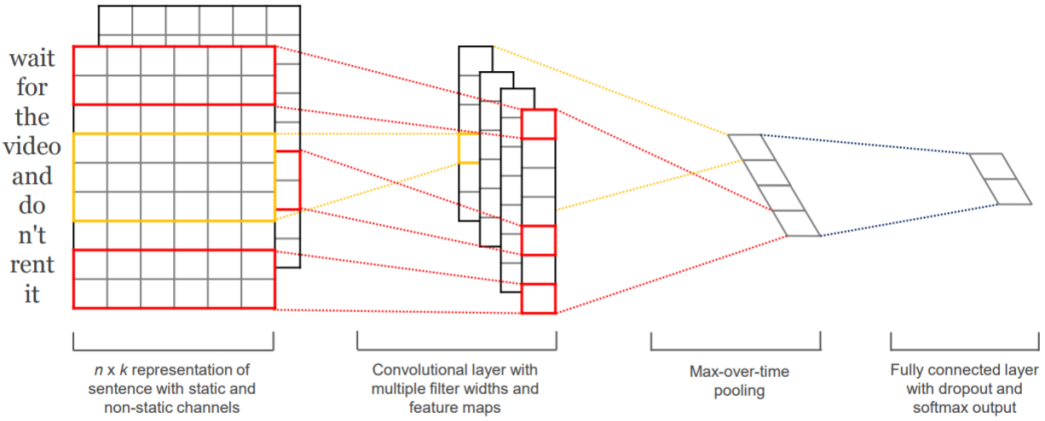

Figure 7: Model architecture with two channels for an example sentence.

3.3 Bi-LSTM (Bidirectional Long Short-term Memory)

Original LSTM [3] (Long Short-term Memory) previously published at 1997 to solve problem which called back-propagation through time. To be more clear, learning to store information over extended time intervals by recurrent back-propagation takes a very long time, mostly because of insufficient, decaying error back-flow.

A Bidirectional LSTM, is a sequence processing model that consists of two LSTM, one taking the input in a forward direction, and the other in a backwards direction. Bi-LSTM effectively increase the amount of information available to the network, improving the context available to the algorithm. For further information go through the original LSTM paper.

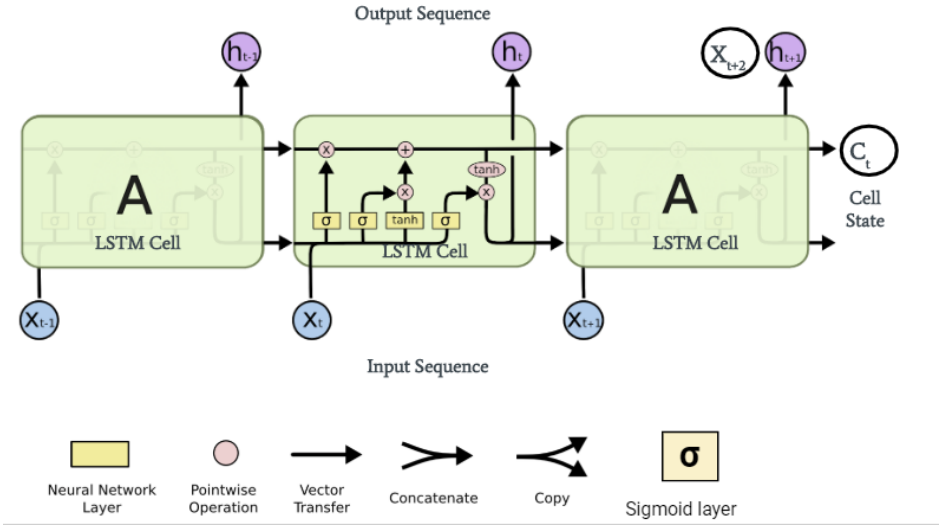

Figure 8: The LSTM cell architecture.

3.4 Fully Connected Layer

Fully Connected layers in a neural networks are those layers where all the inputs from one layer are connected to every activation unit of the next layer. In most popular machine learning models, the last few layers are full connected layers which compiles the data extracted by previous layers to form the final output.

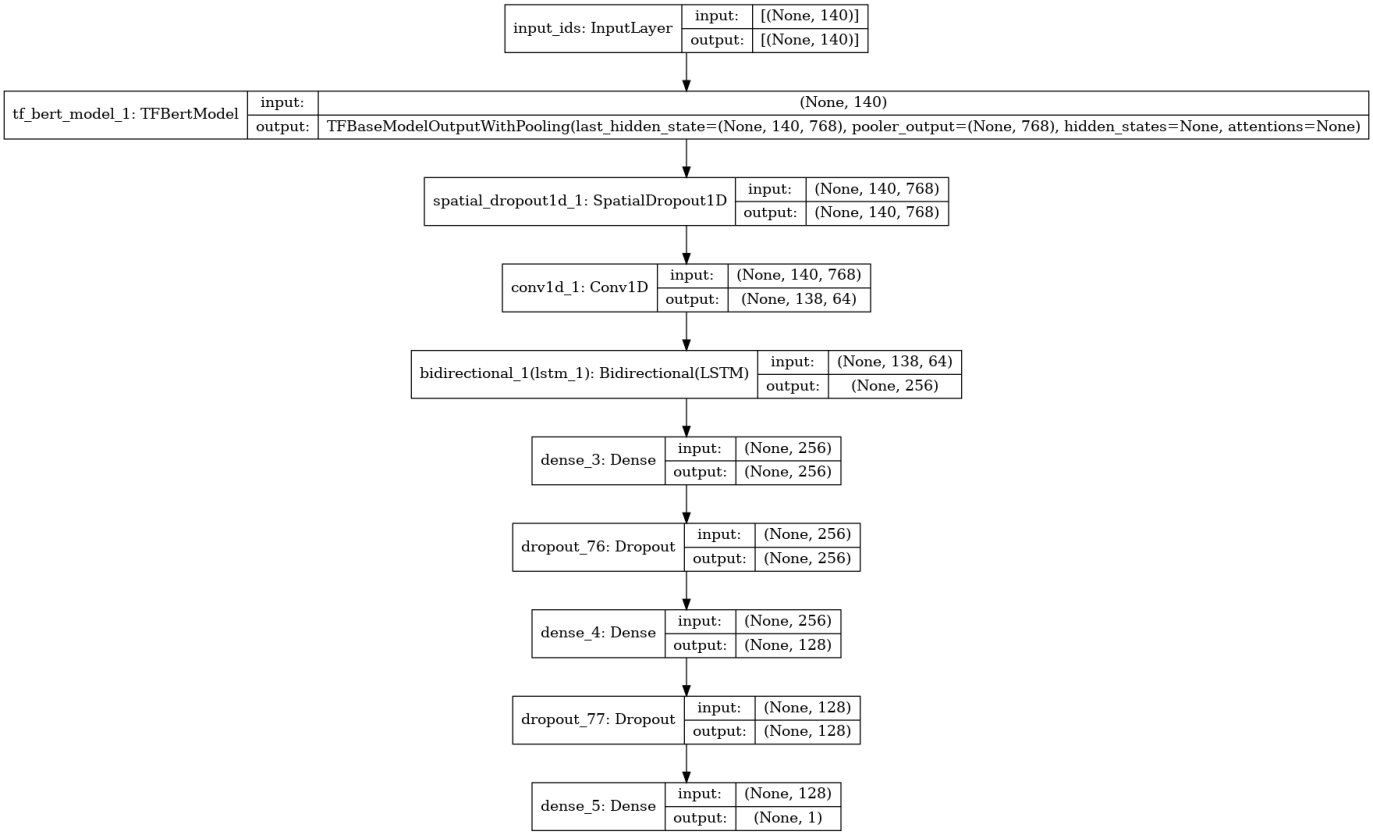

Figure 9: The final model architecture with hyperparameters.

4 Information Retrieval

Twint is an, open source, advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API. Twint utilizes Twitter's search operators to let us scrape Tweets from specific users, scrape Tweets relating to certain topics, hashtags & trends, or sort out sensitive information from Tweets like e-mail and phone numbers. I find this very useful, and we can get really creative with it. The scraped data includes 36 different features. However, we will only consider the "tweet" text feature to evaluate our model. See in table 1. Find performance on random tweets 4

Table 1: Text examples from "tweet" feature for "GRAMMYs" keyword.

Example	Text
1	@hushmirrorball I think she will do something like this in the build up to the Grammys for Evermore. I'm afraid I do indeed need your closure, blondie. :)
2	@RecordingAcad Waiting for metal bands since I saw thisss, me such a dreamer
3	i gotta respect lady gaga for winning last year grammys cause i liked the reaction from #that fandom sooo imma say lady gaga
4	Lana Del Rey getting her Album Of The Year award at the #GRAMMYs (2020) https://t.co/tVGJSf1ukB
5	Krept really disrespected a whole nation who he thinks is a underground something on stage. When this continues how can we bring Grammys to the motherland.
6	Checkout Tunechi's 'drop the world' performance at the 2010 Grammys with Travis Barker.

5 Results and Observations

I will be providing an analysis of mistakes and performance that may shed light on further improvement of the model. Moreover, Twitter scraping library which name is Twint is used to validate our trained model on completely random tweets that searched by keywords such as "GRAMMYs", "elonmusk", and "alternativerock".

5.1 Error Analysis

Table 2 shows six text samples, including three positive and three negative ones, which are misclassified by the proposed BERT-CNN-BiLSTM-FC model. For the first three samples that are marked as positive, none of them are describing obvious meaning. Sample 1 can be detected easily by human because we have sense of extention of 'o' in word "work" most likely gives positive meaning, however, it depends on the flow of the sentence. Therefore, it is challenging to understand difference by machine without enough sentence length. Sample 2 has unclear meaning. Although we might be happy to cat will be fed, cat is lost, and gives sense of negativity. Sample 3 has both emotion in the sentence. Therefore, this kind of situations most likely depend on the dataset we trained the model. Sample 4 is hard to seperate as negative. The length of sentence not giving enough emotion. Sample 5 does not include any negativity. Most likely labelled wrong. Sample 6 true label is wrong again.

Table 2: Examples of misclassified samples. A "+" sign indicates a positive, and a "-" sign indicates a negative sample.

ID	Text	Label	Predicted
1	No wooooook tomorrrow&tuesday	+	-
2	On schedule now to soft launch tomorrow evening - already have first customer - lost cat	+	-
3	@disil429 Fine. Kinders coughing again, so not a lot of sleep. Weather's good, though, that's the main thing	+	-
4	Who else misses Parade the day incredibly much..	-	+
5	can some one pleease tell me how to work out the total surface area of a square pyramid?	-	+
6	@kpsomotragos I have it on good authority that we'll be getting it around 3am.	-	+

5.2 Performance Analysis

I also conducted experiments to evaluate a set of models, and present a performance comparison of all evaluated models in table 3. The set of models CNN, BiLSTM, CNN-BiLSTM, and CNN-BiLSTM-FC forms an ablation study, from which we can evaluate the performance of each individual module and the combined versions. It can be seen that the pure CNN model performs the worst, however, a single-layer CNN can learn many contextual information using BERT embeddings. CNN-BiLSTM present an obvious improvement. My final model, BiLSTM-CNN-FC tops every other model, showing its power to combine the strength of each individual building block. Another observation is that even though the its accuracy is high, I see pure BiLSTM model does not perform robust results in terms of hard cases. Meanwhile, I used Fully Connected Layers to decrease unsteadiness of results in hard cases.

Table 3: A performance comparison of models.

Model	**Precision**	**Recall**	**F1 Score**
BERT-CNN	83.95	85.06	84.55
BERT-BiLSTM	83.81	86.06	84.92
BERT-CNN-BiLSTM	85.61	85.50	85.71
BERT-CNN-BiLSTM-FC	85.75	85.87	85.81

5.3 Prediction on Scraped Data

In prediction phase I used threshold which depends on the prediction score. Label displayed in rules that negative if the score below 0.4, neutral between 0.4 and 0.7, and positive above the 0.7. The results show that totally random tweet sentences can be predicted using the offered scraping library. The model is quite convenient to predict most of the sentences which include any sense. By using sentiment results from random sentences, we can easily observe that the contextual structure and generality is an important advantage of the proposed model, and allows us for further improvement.

Table 4: Samples of "grammys" and "elonmusk" keywords. "ID" refers to keyword.

ID	**Text**	**Prediction**	**Score**
grammys	@peppercqueen @LManoyban @RecordingAcad Are you talking about Lisa?	POSITIVE	0.74
grammys	@RecordingAcad @shattawalegh so sad	NEGATIVE	0.003
grammys	This was during the most recent Grammys	NEUTRAL	0.54
elonmusk	@zalqt1 @elonmusk Not really designed for the winter. The door handles and windows were frozen shut.	NEGATIVE	0.20
elonmusk	@WholeMarsBlog @elonmusk Wow @elonmusk safety competition using FSD: us vs world. It could be interesting if we ever get the same version of FSD going everywhere. When supervised AP safety is proven 20x safer the take rates globally will mushroom up and this could actually be fun.	POSITIVE	0.87
elonmusk	@DERFISCH16 @subcon_deez_nut @elonmusk Ok hob i ned danke google	NEUTRAL	0.58

6 Summarize

Sentiment analysis is highly related to people's daily lives, and recent years have seen more research efforts dedicating to this field. Research on sentiment prediction helps augment people's awareness, and improve the mechanism of a service. This paper investigates a novel model for sentiment prediction using a dirty tweets. My model, BERT-BiLSTM-CNN-FC extracts high-quality linguistic features for sentiment prediction. Although the proposed model is trained and validated on an English dataset, it can be applied to datasets in other languages.

References

[1] Jacob Devlin et al. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *CoRR* abs/1810.04805 (2018). arXiv: 1810.04805. URL: http://arxiv.org/abs/1810.04805.

[2] Alec Go, Richa Bhayani, and Lei Huang. *Twitter Sentiment Classification using Distant Supervision*. 2009. URL: http://help.sentiment140.com/home.

[3] Sepp Hochreiter and Jürgen Schmidhuber. "Long Short-term Memory". In: *Neural computation* 9 (Dec. 1997), pp. 1735–80. DOI: 10.1162/neco.1997.9.8.1735.

[4] Yoon Kim. "Convolutional Neural Networks for Sentence Classification". In: *CoRR* abs/1408.5882 (2014). arXiv: 1408.5882. URL: http://arxiv.org/abs/1408.5882.

[5] Ashish Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.