

## BACHELOR

### Adversarial datasets through sentence length and conjunctions

Sharifzadeh, Ralph A.

*Award date:*  
2022

[Link to publication](#)

#### **Disclaimer**

This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

#### **General rights**

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

**Department of Mathematics and  
Computer Science**

Postbus 513, 5600 MB Eindhoven  
The Netherlands  
www.tue.nl

**Author**  
Ralph Sharifzadeh

**Responsible Lecturer**  
Cassio Polpo De Campos

**Date**  
June 27, 2022

## **Bachelor End Project**

Adversarial datasets through sentence length and conjunctions: how vulnerable are NLP models to the augmentation of long and conjunctive hypotheses?

Ralph Sharifzadeh  
r.a.sharifzadeh@student.tue.nl

# Table of contents

Title  
Bachelor End Project

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Research question</b>	<b>3</b>
<b>3</b>	<b>Previous work</b>	<b>4</b>
<b>4</b>	<b>Methodology</b>	<b>5</b>
4.1	Proposed approach . . . . .	5
4.2	DistilBERT . . . . .	5
4.3	Description of data . . . . .	5
4.3.1	Subsets of the dataset . . . . .	7
4.4	Sentence altering method . . . . .	8
4.5	Validation of new subsets . . . . .	11
<b>5</b>	<b>Results</b>	<b>12</b>
5.1	Preliminary analysis original subsets . . . . .	12
5.1.1	Length subset . . . . .	13
5.1.2	Conjunction subset . . . . .	15
5.1.3	Combination subset . . . . .	16
5.2	Augmented subsets analysis . . . . .	16
5.2.1	Augmented length subset . . . . .	17
5.2.2	Augmented conjunction subset . . . . .	19
5.2.3	Augmented combination subset . . . . .	20
5.3	Comparison between unedited and augmented subsets . . . . .	22
<b>6</b>	<b>Discussion</b>	<b>23</b>
6.1	Future work . . . . .	24
<b>7</b>	<b>Conclusion</b>	<b>25</b>

# Table of contents

---

Title  
Bachelor End Project

**References**

**26**

## **Abstract**

In this thesis there will be an attempt to break NLP model called DistilBERT. The motivation for doing this is the fact that fake news becomes more and more present in the current day and age. Social media platforms are over flooded with fake news articles and it can be difficult to know whether something is fake news or not. The experiment will be carried out by augmenting several subsets of the MNLI training data. The data is divided into a subset with long hypotheses, hypotheses with conjunctions, and the combination of the two. The research question that is answered after analyzing the results of the subsets goes as follows: Adversarial datasets through sentence length and conjunctions: how vulnerable are NLP models to the augmentation of long and conjunctive hypotheses? It is shown that the DistilBERT model is vulnerable to the augmentations made in this context of the battle against fake news.

Keywords: *NLP, NLI, Fake news, Adversarial datasets, BERT, DistilBERT, TextAttack, Sentence Similarity*

# 1 Introduction

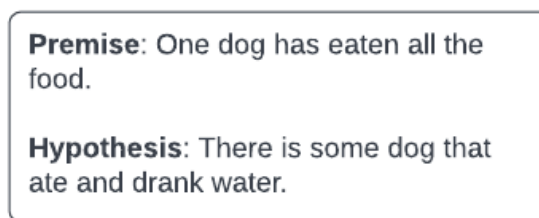
Natural Language Inference (NLI) is one of the main downstream tasks in Natural Language Processing (NLP) and its research (Bowman et al., 2016) concerns premises and hypotheses that are compared to each other to make out whether they are contradictions, entailments, or neutral combinations (Belinkov et al., 2019). However, changes in these sentences have shown that models can be broken relatively, maybe even disturbingly, easily (Glockner, Shwartz, & Goldberg, 2018). The fact that changes, often only small, lead to these NLP models breaking could mean that the content on which a model like BERT (Devlin, Chang, Lee, & Toutanova, 2019) is pre-trained lacks insights into certain patterns. Pre-training is the act of training an algorithm on one task so that it can make parameters to do other tasks. A lack of this results in the fact that model does not learn patterns. Finding and creating adversarial examples to break these models and show these gaps in the models understanding (Minervini & Riedel, 2018; Goodfellow, Shlens, & Szegedy, 2014; Szegedy et al., 2013) is important to find out where and why models make the mistakes that they make.

An interesting area in which this is important is regarding fake news. A pertinent example is the comparison between titles of articles to see whether they match or not (Yang, Niven, & Kao, 2019). The checking is originally done by humans themselves but this takes immense amounts of time. In the current day and age, there should be more automation involved with this process (Oshikawa, Qian, & Wang, 2018). There are lots of technical challenges involved with fake news detection. The continuous growth of social media platforms makes fake news spread faster and faster. A problem with this is that many children and young adults are heavy users of these platforms (Álvaro Figueira & Oliveira, 2017). For them, it is more difficult to realize that a header or article is in fact fake news like in figure 1.1. The fully technical aspect to it is that it is difficult to know what the actual fact is. Automating finding facts in a huge landscapes of millions of reports is just not possible yet. It is therefore important to think of more adversarial examples in order to improve NLP algorithms and models. This will slowly but surely overcome some of the challenges regarding fake news.



Figure 1.1: An example of a fake news article on Facebook

With this in mind, the goal of the paper is to add to the already done research into NLI and its influence in fake news detection in modern times. The focus will be on changing sentence-pairs from an NLI dataset. Specifically, the sentences that will be changed are the hypotheses in order to see how the performance of NLP models changes with respect to the sentence length in addition to the presence of coordinating conjunctions. These two separate aspects and the combination of the two is chosen because pieces of articles contain long sentences which leads to the usage on conjunctions in order to form longer sentences. Titles of fake news articles, as mentioned earlier, can be lengthy as well. Therefore, they are more likely to contain conjunctions in them. It is thus important to look specifically at the sentence length. Next to this, models can have difficulties with conjunctions if they are present in a sentence (Saha, Nie, & Bansal, 2020). An example of both of these aspects is shown in figure 1.2.



**Premise:** One dog has eaten all the food.

**Hypothesis:** There is some dog that ate and drank water.

Figure 1.2: Example with a longer hypothesis and conjunction in it

## 2 Research question

The problem addressed in the introduction made it clear that there will be a look at specifically longer hypotheses and those with conjunctions in them. Next to that, the combination of these two is also of interest to see what type of content it is that overlaps. The augmented subsets have as a goal to break the NLP model. This means that the research questions asked should relate to these subsets and the goal to break the model. The question risen is whether NLP models are vulnerable to changes in augmented sentences and this leads to the following main research question that is proposed: **Adversarial datasets through sentence length and conjunctions: how vulnerable are NLP models to the augmentation of long and conjunctive hypotheses?**

The main research question has been introduced but it is beneficial to add multiple sub-questions to guide the process of the methodology and analysis. This paper will try to answer the main research question via the sub-questions below:

- How is the term 'vulnerable' defined in this specific context?
- How often do coordinating conjunctions appear and do they then appear in a hypothesis that is longer than the premise too?
- Is there a baseline negative impact on the model performance when the hypothesis is longer than the premise itself?

The first sub-question is important especially for later on in the experiments when there will be comparisons and results. How do we decide whether these results and intermediate results show vulnerability or not? This will be addressed later on. Secondly, we want to know if there are sentences, hypotheses in particular, that have conjunctions in them to start with. There might even be overlap between the longer hypotheses and these hypotheses with the conjunctions in them. The aim is to look at whether they are present and in what numbers. Lastly, it might be beneficial to see if longer hypotheses examples without any perturbations of any kind perform worse on the model than other rows do.



### 3 Previous work

There are multiple pieces of work that concern the same problems or something related to this research topic. For example, a similar research paper is the one from Li et al. (2020) in which the authors create adversarial examples, in their case in the Chinese language. They make their own model in order to find the best replacement for a word to create essentially the best adversarial example. This is a great development in the field and for that specific language however it does not touch upon the NLI part of the field and the research does not incorporate sentence-pair classification of any kind. It leads to checking the perturbed sentences to see if they make sense next to whether they seem like convincing examples.

Another paper that is mentioned as previous work is the paper about PAWS (Zhang, Baldrige, & He, 2019). The first thing about the paper is that it is a highly linguistically oriented paper. In the paper they create a new dataset named PAWS which contains sentence-pairs that are either paraphrases or not but all the sentence pairs have a high bag-of-words (BOW) overlap. Here again the sentence-pair aspect is not captured and the sentences are not completely new. This is of course because of the high BOW overlap meaning that the goal is not to change many words but rather the opposite.

A main piece in the realm of relevant work is a paper that evaluates a big amount of NLP tasks, one of them being the task of testing out sentences with coordinating conjunctions (Kim et al., 2019). However, they test whether two elements before and after a coordinating conjunction are a contradiction, entailment or neutral combination. There is no comparison between premises and hypotheses. There is another paper which is about NLI and sentence-pairs. In the paper the authors look at whether the label can be inferred from the hypothesis only (Gururangan et al., 2018). Even more, there is a short mentioning of the sentence length in the paper describing observations of correlation between this and the label itself. The main point however is the analysis of certain keywords that are associated with certain labels. The paper will serve as a reference later on for this research project as well to for example compare results or see if there is a correlation between certain words and labels. Gururangan et al. (2018) does not incorporate the invention of newly created adversarial examples in any way and the coordinating conjunctions are also not addressed in the paper.

## 4 Methodology

The methodology used in this research project will be explained in this section as it will provide the necessary information about the dataset, the way the dataset will be altered, and why and how the newly created dataset will be validated. After revising the previous work done regarding this topic, like the work from Kim et al. (2019), it is important to start off with describing the approach that will be taken for the rest of the paper. Then, a description of the chosen dataset will be given to highlight the data that will be used. This dataset is sourced from the datasets library of Hugging Face and are also imported for the programming parts using the Datasets package for Python.

### 4.1 Proposed approach

In order to possibly break a model we need to augment the dataset in such a way that will give us the subsets we want to test for. Thus, the approach taken will be done on sentence level. This means that the perturbations the dataset will receive consider changing parts of the sentence, specifically the hypothesis of the sentence-pairs, rather than changing words like negating a word. Furthermore, it will become clear that the manipulations are not limited to a single word meaning there might be more than one word changed in the sentence. In this section, the methods that are used for the approach are further clarified. The specific framework is presented and the process is explained and shown visually as well. This process is to go from just a subset of the original dataset to the eventual edited subsets that will be used in the analysis to try and break the NLP model.

### 4.2 DistilBERT

It is important to note that the performances and analyses are recorded with the NLP model called DistilBERT. DistilBERT is a distilled version, as the name suggests, of the BERT model (Sanh et al., 2019). This new pre-trained model is capable of being used for the wide range of tasks that its larger counterpart can also be used for. However, DistilBERT is smaller, faster, and lighter (Sanh et al., 2019) making it a good option to use for our experiments. It is therefore possible to use our complete subsets to be run through the model and get the metrics we need.

### 4.3 Description of data

First and foremost, the dataset of the study is the Multi-Genre Natural Language Inference (MNLI) dataset from Williams et al. (2018). The dataset consists of sentence pairs which are regarding several different genres. Written and spoken text is included in this dataset as well to make up a wide range of sentence-pairs. This is also the dataset that will be altered for the research purposes later on. Next up, the goal is to establish more precisely the content of this dataset. The dataset consists out of 392.702 rows that come from the training set of the MNLI dataset. The rows consist out of three dimensions or columns: the premise, the hypothesis, and the label.

The premise column is a string variable and serves as the basis piece of text on which the hypothesis is evaluated. This hypothesis is often, however not always, a shorter string which often holds information related to the content of the premise. In figure 4.1 you can see the distribution of premises and hypotheses according to their length. Indeed it shows the tail of the premises meaning that there are more longer premises than hypotheses. However, within the overlap of the hypotheses and premises we can also see that there are cases with a longer hypothesis.

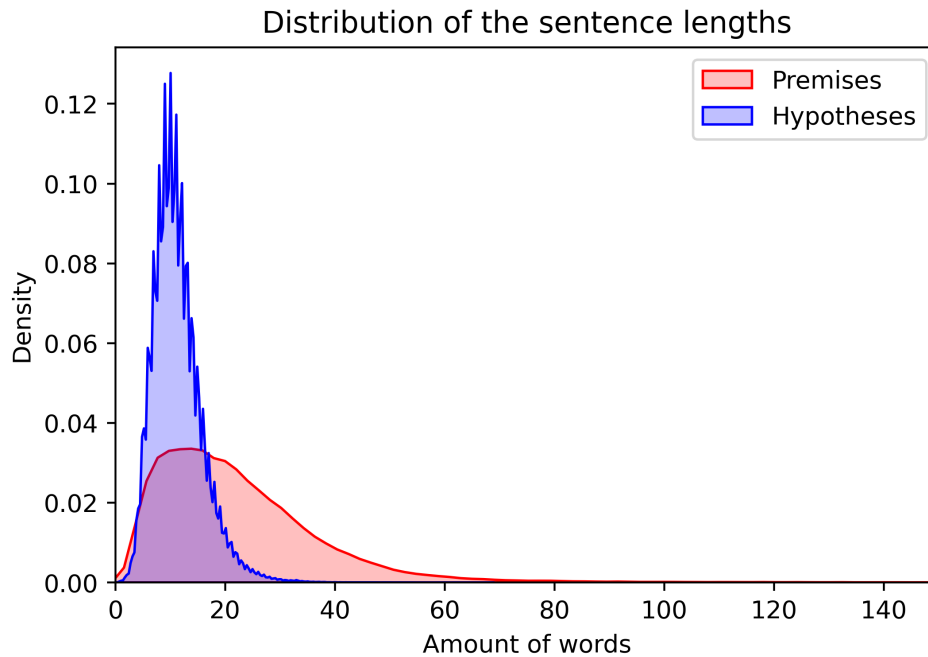


Figure 4.1: Distribution of the sentence lengths

Whether this hypothesis entails or contradicts the premise is stored in the final variable, the label. The label shows the conclusion of whether the two are entailments, contradictions or neutral in comparison. An example of each possible label in the dataset can be seen in table 4.2.

Premise	Hypothesis	Label
'You will go back to them.'	'You'll find your way back to them.'	'entailment'
'VA announced its first Carey Award in 1992.'	'Carey Awards was created in 2009.'	'contradiction'
'It seems to me very simple.'	'It didn't have many issues.'	'neutral'

Figure 4.2: Examples from the dataset

There are a few remarks to make regarding the table above. First of all, the very first row is an example case where the hypothesis is longer than the accompanying premise. This will be one of the subsets talked about in the methodology section. Another interesting remark to make ties in with another piece of work that was also mentioned in the previous work section. Gururangan et al. (2018) mentioned in their paper that neutral hypotheses tend to be long and entailment hypotheses short. The example on the first row we see here with the longer hypothesis however shows the entailment label. This could be for several reasons, for example the fact that even

though the hypothesis is longer than the premise it is not long in tokens. It could also just be an example that deviates from the overall findings from Gururangan et al. (2018).

### 4.3.1 Subsets of the dataset

In the previous section, some insights were given about the content of the dataset together with basic summary statistics. This previous overlook however summarizes the complete dataset without going deeper into the content. In the light of the research topic it is more interesting to also look at some specific cases in the dataset. Firstly, the cases which have a longer hypothesis than premise are isolated from the rest. In figure 4.2, the first example is actually one of these cases since the hypothesis is longer than the premise. Even though this is not the most abundant type of premise and hypothesis in the dataset, it is definitely not the only example. In total there are 62,175 rows which show have a longer hypothesis. This comes down to around 16% of the original dataset being of this origin. The statistics for this subset of the dataset can be seen in the figure 4.3 below.

Label	Amount of rows	Percentual portion
'neutral'	28.463	46%
'entailment'	16.341	26%
'contradiction'	17.371	28%

Figure 4.3: Statistics for the subset with longer hypotheses

There is another subset that is explored and used for later analysis. The basis for this subset of the dataset follows from a more linguistic point of view, namely it considers the coordinating conjunctions in the hypothesis. The definition can be given as the following: *a connective or connecting particle with the special function of joining together sentences, clauses, phrases or words* (Gleitman, 1965). These connectives words are extremely important in the formation of sentences which is also mentioned in the introduction regarding the motivation of why this experiment is done. There are a total of seven conjunctions that fall in this category and the definition but two of these will be chosen as the focus points: the words 'and' and 'or'. The MNLI dataset contains a total of 61.207 rows which have at least one of these two words in their hypothesis. The way the long hypotheses were compared via the label distribution is also applied here. The biggest group however for this subset is not the label 'neutral' but rather 'entailment'. It is safe to say that overall the labels are quite equally distributed. The numbers can be seen in figure 4.4.

Label	Amount of rows	Percentual portion
'neutral'	20.378	33%
'entailment'	21.034	35%
'contradiction'	19.795	32%

Figure 4.4: Statistics for the subset with coordinating conjunctions

So far two subsets have been established and there is one last one that is constructed from the two established ones. Thus, this is going to be the intersection between these two. The amount of rows that have this shared set of characteristics is 11.194. This is only a little subset of the original dataset at around 3% of the total amount of rows. Nonetheless, the same summary statistics for this subset are given in figure 4.5 and the 'neutral' group occupies almost half of all the rows in this category. It looks very similar to the distribution of the labels for the subset with the rows that have a longer hypothesis than the premise.

Label	Amount of rows	Percentual portion
'neutral'	5.493	49%
'entailment'	2.605	23%
'contradiction'	3.096	28%

Figure 4.5: Statistics for the subset with both a longer hypothesis and a coordinating conjunction

## 4.4 Sentence altering method

The subsets will be perturbed using the TextAttack framework (Morris et al., 2020) to augment the selected sentence-pairs. The TextAttack framework is a relatively new framework that originates from 2020 and allows for convenient construction of new, augmented examples from a dataset. It enables other processes as well such as using it to train a model itself. In figure 4.8 the process is visually described. Each subset will be subject to perturbation by TextAttack.

First, the hypothesis of the sentence-pairs are isolated from that particular row. By doing this, the individual sentence can be altered. To understand the perturbations that will be done it is important to clarify what TextAttack needs as input and what it will give as output. This is because from TextAttack, a specific function is used to create the adversarial examples called TextAttack Augment. The augmentation functionality of the framework provides several ways to control how you want to perturb sentences. This will be important since the goal is to perturb the target sentences in specific ways. Before describing how the specific settings are chosen it is important to look at some characteristics of the augmentation functionality. The specifications for this function can be seen below in figure 4.6.

Parameter	Description
Input file	The csv file which holds the dataset to be augmented
Output file	The csv file to which the new adversarial examples are saved
Input column	The column of the input file which will be augmented
Recipe	This defines the nature of the changes that will be made
Pct words to swap	The percentage of words in the sentence to be swapped
Transformations per example	The amount of newly generated examples per original input
Exclude original	If included, the original rows will be excluded

Figure 4.6: Input parameters for the Augment function of TextAttack

The application of this framework for our research heavily impacts the way that the parameters above are chosen. One of the most important aspects and maybe the most important one is the recipe is used. We are namely not interested in the insertion or deletion of words to the sentence we augment. This is because the original length of the hypothesis should be maintained. Suppose that we do allow for insertions and especially deletions. This could turn a hypothesis that was longer than a premise at first into a shorter hypothesis.

There are two recipes that seem to be a good fit in our case and these are 'wordnet' and 'embedding'. The first recipe, 'wordnet', replaces words by putting in an alternative from the WordNet Thesaurus. It chooses a synonym of the word to do this. The other recipe mentioned, 'embedding', augments text by replacing words with neighbors in the counter-fitted embedding space, with a constraint to ensure their cosine similarity is at least 0.8 (*Augmenter Recipes CommandLine Use*, 2020). The cosine similarity says something about how close words are in a predefined space. The cosine of an angle between two lines, or vectors, to two words describes whether they face the same way and therefore are similar. A visual representation can be seen below in figure 4.7.

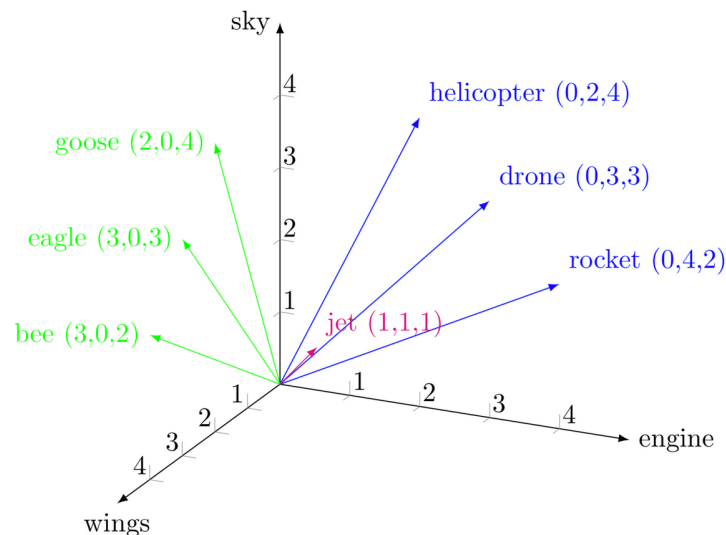


Figure 4.7: Visual example of word embeddings

Both of these can result in the type of adversarial examples that are desired. Looking at the other required parameters we move on to the percentage of words to be swapped. It is not the goal to change the entire sentence and there is the risk that a high percentage here would result into the generated examples being completely worthless. Especially when we think about longer sentences this comes into play. A percentage which is too high will cause more words to change if the sentence is indeed long. The result could possibly be that the new augmented sentence has so many words changed that it loses its meaning altogether. Therefore, this percentage is kept at a level which is around 10%. This value is chosen to generate the examples. Another deliberate choice is to maintain the total amount of rows in the subset. The sample sizes of the subset and the perturbed subset will be equal in this case as the original rows will be excluded from the new perturbed version. Below in figure 4.8 you can see the aforementioned visualization describing the entire process from beginning to end.

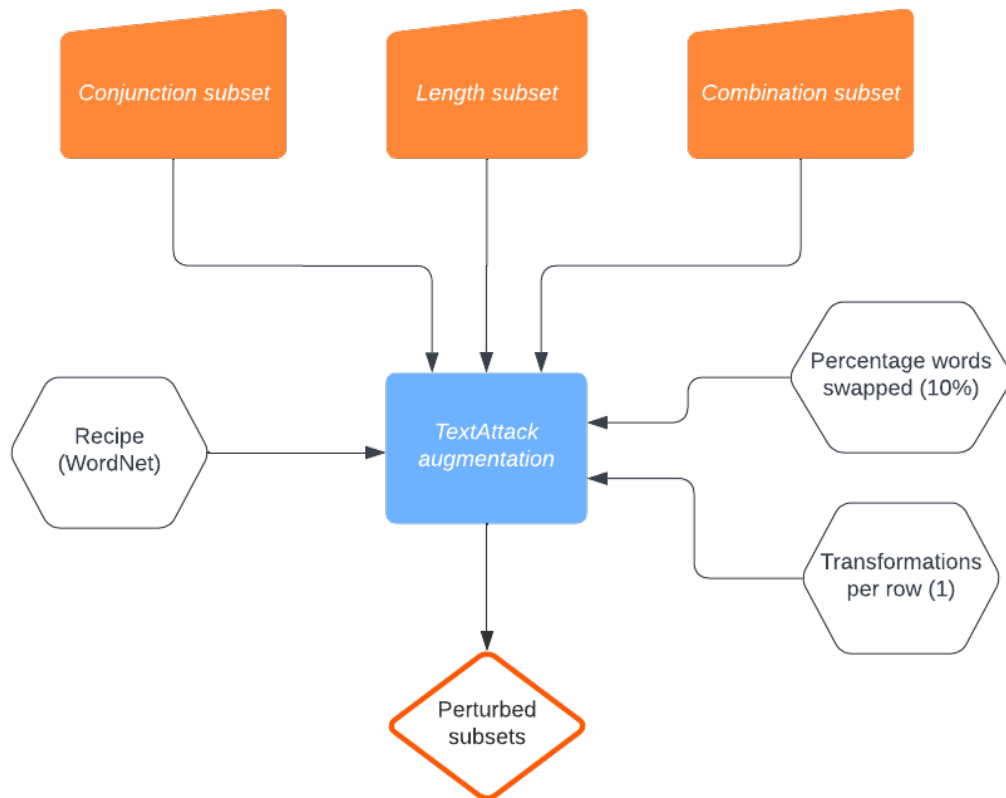


Figure 4.8: Process visualization for perturbation with TextAttack

The complete process shows that the new datasets should be constructed in such a way that they can be compared to their respective unedited counterpart. Next to that, they should be able to be compared to the other subsets in general. This is important for the analysis of the datasets in a later chapter. An example row of this output can be seen in figure 4.9, which is a sentence from the subset with only the coordinating conjunction sentence-pairs. For clarification of the example, the original sentence is also added to display the difference between the two. However, the original sentence is excluded from the perturbed dataset itself. Also, it is important to note that this is an example generated with 'wordnet' as the recipe for the augmentation with TextAttack.

Form	Sentence
Original	'The kiosks offer English only information and tours.'
Perturbed	'The kiosks provide English only advice and tours.'

Figure 4.9: Example of a perturbed hypothesis

## 4.5 Validation of new subsets

With the operations performed in earlier steps it is evident that multiple new subsets are created. It is not completely newly generated as is done by other papers mentioned in the section regarding previous work. The way to see how the augmentations turned out is done via a threshold regarding the sentence similarity. Sentence similarity is used to see whether two sentences are similar and how similar they are (Wang et al., 2016). It will be needed at the very end of the augmenting process. Augmentations will inevitably lead to some hypotheses being wrong. By using a sentence similarity score, it is possible to throw out certain rows that are not similar to the original hypothesis anymore.

In table 4.10 you can see an example where the sentence doesn't make sense anymore after the perturbation. This is a sentence from the subset containing a conjunction and a longer hypothesis so it may seem like nothing much changed. However, the word that changed messes up the sentence. If a person sees the perturbed sentence only, they might think that there is some literal visualization involved because of the word choice. But in fact the real meaning is that it is looked upon as a miracle. In another example, 'United States' turned into 'United says' because it replaces the word 'states' with a literal synonym. Naturally, this did decrease the amount of examples that can be used for the label classification of these subsets.

Original	He created both Hindu and Buddhist temples and statues what was seen as a miracle.
Perturbed	He created both Hindu and Buddhist temples and statues what was visualize as a miracle.

Figure 4.10: Example of a perturbed mistake

These types of sentences are not wanted in the remainder of the analyses. Therefore, we use the similarity score to filter the rows. This score is based on the cosine similarity score that was shown in figure 4.7. The texts are put into an embedding and two different sentences are compared to result in the scores. In this case, the newly created hypotheses are compared to the original hypotheses. The score 0.6 is chosen since this throws out those sentences that don't make sense anymore. Below in table 4.11 you can see an example sentence like this with an insufficient cosine similarity score. Even though only one word changed, the cosine similarity score is 0.54 so it is being cut off.

Original	Understand, I have been here for only four hours.
Perturbed	Understand, iodine have been here for only four.

Figure 4.11: A perturbed hypothesis that does not have the original meaning anymore

For the remainder of the analysis the subsets have been filtered using this cosine similarity threshold of 0.6.



## 5 Results

First of all, the original subsets are investigated on how the perturbations turned out after the previous phase. The importance lies in the question whether the perturbations turned out to ruin the examples in such a way that they cannot be used anymore. With sentence similarity used as a metric to maintain quality, this is an interesting question. As mentioned before, the DistilBERT model is used to classify the sentence-pairs with a new predicted label.

### 5.1 Preliminary analysis original subsets

The three subsets (length, conjunction, and the combination) are classified by the model that was chosen beforehand. They are fully processed by the model in order to see how their predicted labels compare to the original labels that they had attached. In figure 5.1 the results per subset are displayed.

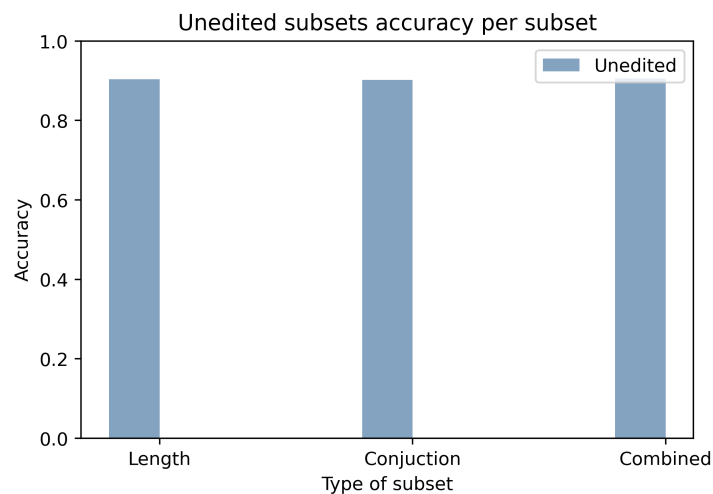


Figure 5.1: The accuracy for the three unedited subsets

The values on the unedited training set are quite high for the DistilBERT model. This makes sense since the data is part of the MNLI training set. Even though it is split up into different subsets, it is still the same training data. The differences between the unedited subsets is also very small and almost insignificant to reflect upon. They all fall within the same 1% of each other (90.2%-90.6%). The length subset does not seem to have a big baseline negative impact on the accuracy. This does not say anything about the perturbed version of the subset yet as this hasn't been seen yet. One of the subquestions to the research question regarding this cannot be supported from what is seen so far. Now it's important to look at the individual subsets and see if there are certain specific sentences that the model does not classify well even on the training set itself.

### 5.1.1 Length subset

The length subset will be looked at closer in terms of wrongly classified rows. The length subset shows the failures that it has are mainly on neutral labels that should not be neutral. Gururangan et al. (2018) also stated that there seemed to be some correlation between a long hypotheses and a neutral label. However, we should take into consideration that the biggest percentage of labels was neutral. There hasn't been a shift in label distribution. Therefore it is not correct to say at this point that relatively the neutral label was the one with the highest fail ratio. Below in figure 5.2 you can see the the amount of failures per label.

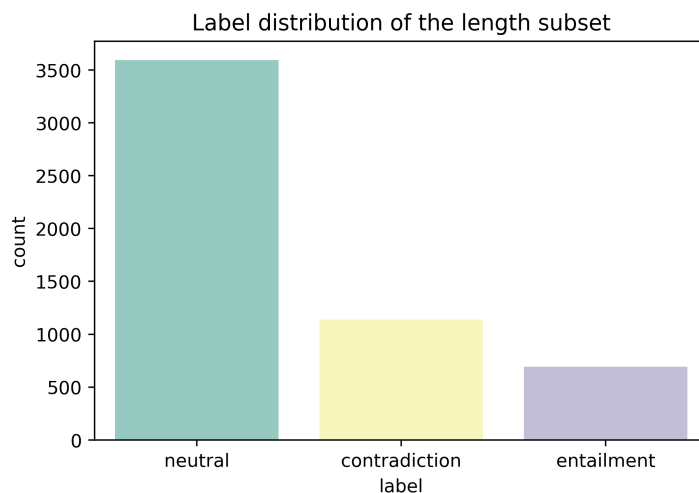


Figure 5.2: The amount of failed rows per label for the length subset

Another interesting look into the subset is to see the sentence length of these failed examples. As discussed before, there didn't seem to be a negative impact by sentence length compared to the other subsets. What we want to check is the sentence length of the failed examples. In figure 5.3 a plot is made similar to figure 4.1. It shows the density plot of both the premises and the hypothesis.

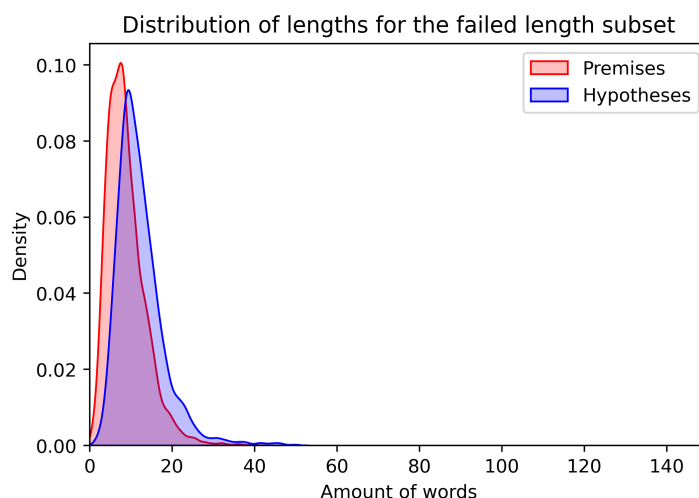


Figure 5.3: The distribution for the word lengths of rows that were wrongly classified. This regards the length subset

Since this is the length subset it makes sense how the plot looks. The hypotheses have to be longer than the premises so the curve of the hypotheses will always be in front of the premises one. However, looking at the axes it does seem like they are extremely long compared to all rows. A way to compare this is to see the plot of all the rows and visually look at the difference between the plots. In figure 5.4 you can see the plot for the whole length. It shows similar numbers on the x-axis so it doesn't seem like the failed examples of the unedited length subset are abnormally long.

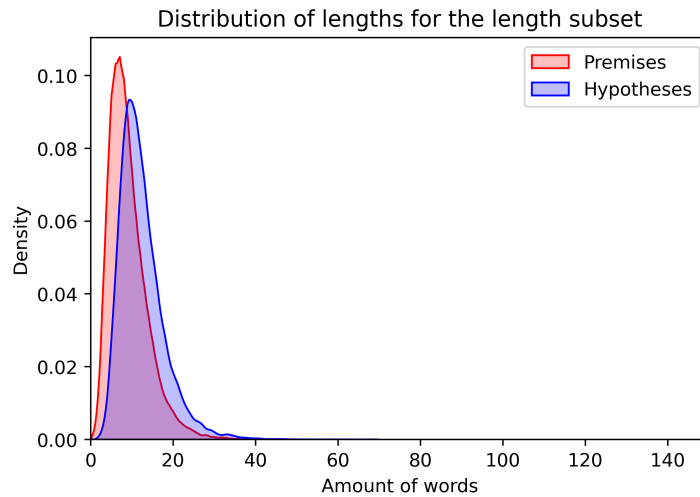


Figure 5.4: The distribution for the word lengths of all rows. This regards the length subset

### 5.1.2 Conjunction subset

The conjunction subset has the lowest accuracy even though in this particular setting 'lowest' is not much lower than the rest. An interesting look into the dataset is to see the most common words in the hypotheses overall. Here we can see an interesting result since a word is missing from the figure that we might expect to be there. It can be seen in figure 5.5 below. The word 'and' is easily the most common word together with some generally common words. The other conjunction, 'or', is missing from this figure.

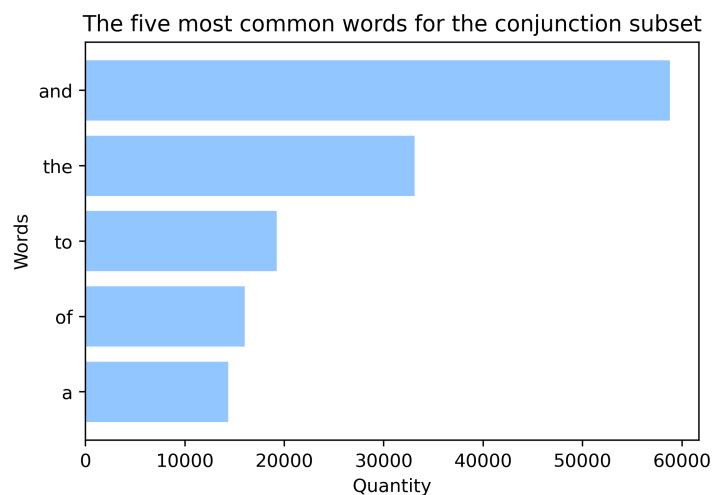


Figure 5.5: The five most common words in the unedited conjunction subset with 'or' missing from this top five

Looking at the failed examples from this subset and their five most common words it is identical to the plot above. The only difference is the range on the x-axis since the size of this failed subset is smaller. The word 'or' is also not in these failed examples and it seems like it generally appears less in the dataset overall. There is another effect that shows up for this subset. It regards the shift of label distribution for the failed examples. For the length subset we did not see a big difference in this distribution but for the conjunction subset that's a whole other story. It can be seen in figure 5.6.

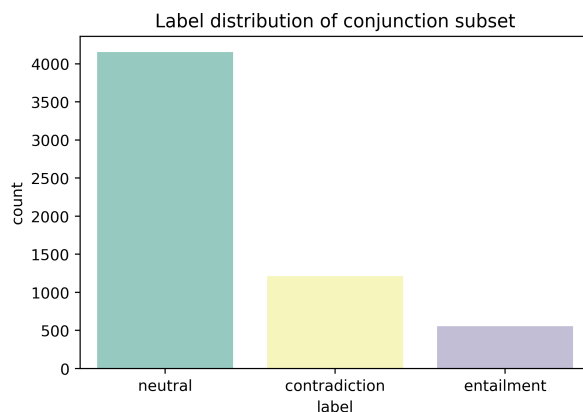


Figure 5.6: The distribution of labels for the conjunction subset. It is the distribution for the rows that were wrongly classified

It is visible that the neutral label has the vast majority out of all failed examples. This is in high contrast to the evenly split in labels the conjunction subset originally has. It seems to be more difficult for the model to classify the neutral labels as neutral correctly. And this is even the case for the unedited subset in this case. An example of one of the wrongly classified rows is seen below in table 5.7. This can be seen as a rather difficult example. The premise has this extra clause in the middle but for the rest it doesn't seem like there are strange things going on with the sentence.

Premise	His scalp, where he had cut away his dark hair, was pale.
Hypothesis	His scalp was shiny and exposed.

Figure 5.7: One example of a wrongly classified row where the original label is neutral

### 5.1.3 Combination subset

The combination subset is addressed in a smaller manner here. It takes over some of the characteristics that are highlighted for the other subsets. The distribution of the sentence length also never allows the premises curve to overtake the hypotheses one like in figure 5.4 or figure 5.3. This would mean a longer premise than a hypothesis. The labels follow the conjunction distribution closer as in figure 5.6. This also aligns with the original label distribution of the combination subset.

## 5.2 Augmented subsets analysis

Now we will look at the augmented subsets. These subsets are again the length, conjunction, and combination subset. However, this time the hypotheses considered unless mentioned otherwise are the newly created ones. Initially the accuracy maintained from the augmented subsets is shown below in figure 5.8.

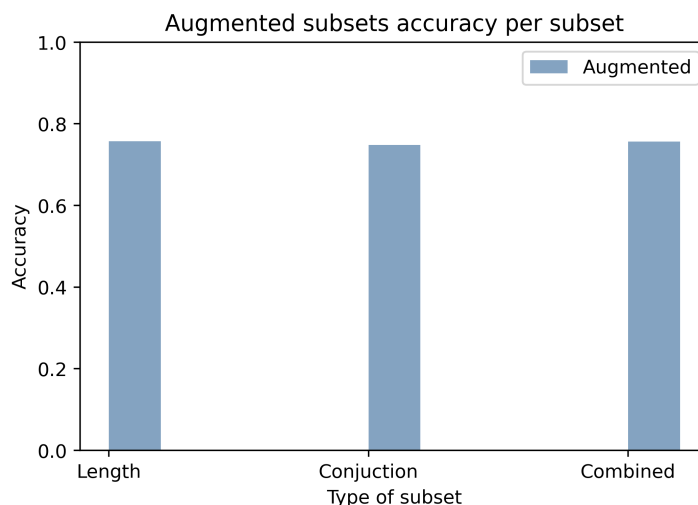


Figure 5.8: The accuracy for the three augmented subsets

The values for all the three subsets are lower for the augmented versions. Between each other they are very close to each other without one having a much worse accuracy than another subset. The augmentations did have impact on the accuracy. With the cosine similarity as the sentence similarity threshold we assume the quality of the subsets to be good. This means that the rows included to reach this accuracy should have a valid augmented hypothesis. The results of the sentence similarity is an interesting metric to look at further per subset too. First to remind what the augmented hypotheses look like, in table 5.9 there is an example.

Original	Which food is that? the green one?
Perturbed	Which meal is that? the green one?

Figure 5.9: One example of an augmented hypothesis where the noun is changed from 'food' to 'meal'

### 5.2.1 Augmented length subset

For the length subset, we will look at the distribution of the similarity scores. It can be deduced from the methodology process that the value of this similarity will not be lower than 0.6. This is because these rows were deleted from the subsets. From the remaining rows it is interesting to see whether the labels influence the similarity score of the row. In other words, is for some label the distribution of the similarity score deviating from the other labels? In figure 5.10 the distributions are plotted with the usage of a violin plot. The y-axis shows the different labels that exist.

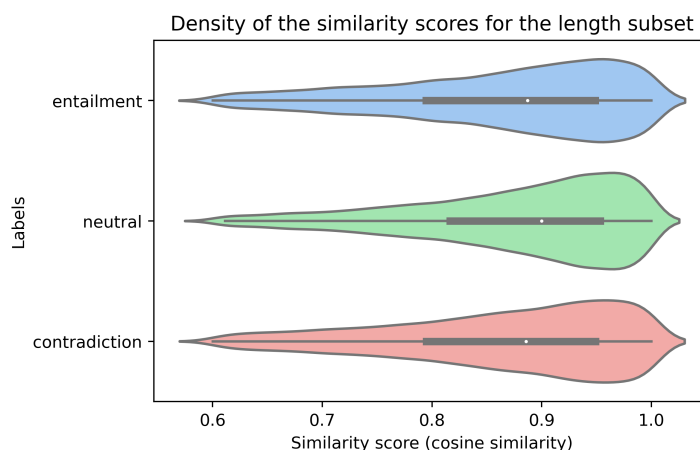


Figure 5.10: The distribution of the similarity score over the three different labels. The similarity score is the computed using the cosine similarity

There are no big differences in the distributions here. The neutral label stand out a little bit since the left tail of the distribution is a bit slimmer than the rest. At the same time it has a bit more bandwidth closer to 1. The fact that all three distributions lean towards the right side of the spectrum makes sense. We get rid of the bad rows and at the same time we only change 10% of the words in each hypothesis. This means that the hypothesis does not change incredibly much while lowering the accuracy of the DistilBERT model. The example in table

5.9 is one of these hypotheses. Only one word is changed here and the similarity score remains high at around 0.93.

Next, we will look at words to often co-occur together in the augmented hypothesis. This is calculated using Pointwise Mutual Information, PMI. The equation for this value is

$$PMI(a, b) = \log\left(\frac{P(a, b)}{P(a)P(b)}\right) \tag{5.1}$$

This equation says something about probability of both events happening at the same time compared to their individual probabilities. In our case,  $a$  and  $b$  are two words. Suppose that the words are completely independent of each other, then the formula will result into a value of 0 due to the logarithm. Now that we know what PMI entails we can see how the hypotheses of the augmented length subset score. In figure 5.11 the top ten word combinations based on the PMI value are shown.

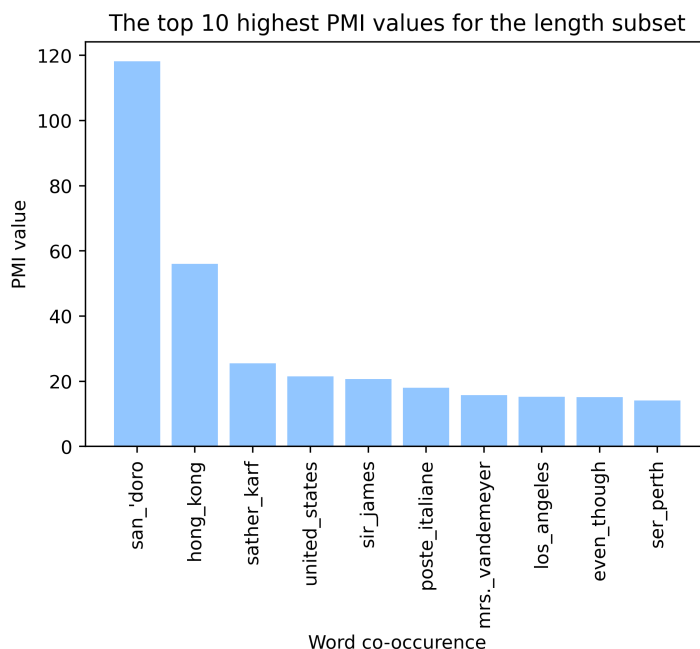


Figure 5.11: The top 10 word combinations based on their PMI values

What is obvious from this plot is that there are certain names of places that show up in this plot. We can see Hong Kong and Los Angeles for example. It makes sense they appear together often with such a high PMI value. Furthermore there are also some names of people that fall in this same category. There might not be many rows containing these names but they are so connected to each other that their PMI value is high. Word combinations past the top ten also show the same type of observations. There are car names that co-occur for example. These are sometimes not even real words and only occur together with the other part of the username. An example from this is the name 'mazda rx7'. Its individual probability is incredibly low whereas the co-occurring probability is high. Below in figure 5.12 there are some more co-occurring words plotted showing some more names that can be seen. There are in this case also some related terms like

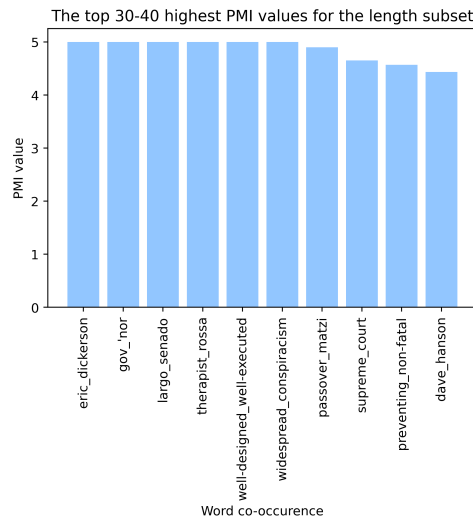


Figure 5.12: Places 30 to 40 of the word combinations based on their PMI values

### 5.2.2 Augmented conjunction subset

The augmented conjunction subset had an extra constraint to it while augmenting. The way the augmenting was done there were no additions or deletions, but only replacements. However, it is vital that for this specific subset some words are also not replaced. This concerns the conjunctions themselves, namely 'and' and 'or'. If the conjunction is removed from the sentence then it does not belong into this subset anymore. The subset was therefore checked for this and all rows still had either one or both of the conjunctions in them.

The subset will be illustrated using a violin plot in the same way as the augmented length subset. It will show the distribution of the similarity scores for this subset. In figure 5.13 you can see the plot. The left tail of each label is almost not there. The bulk of the values lay past the 0.85 point. It is even more extreme than for the augmented length subset seen in the section before. For the rest it also has no values below 0.6 because of the cut-off point set by the sentence similarity threshold. The average similarity score is also very high at 90.5%.

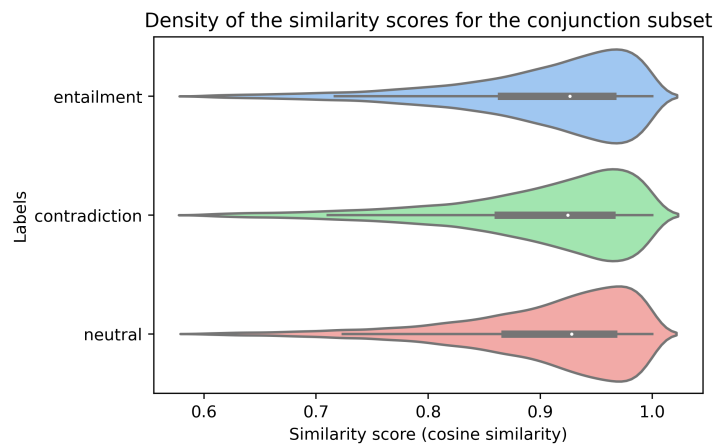


Figure 5.13: The distribution of the similarity score over the three different labels. The similarity score is the computed using the cosine similarity



The distribution of the sentence length might be interesting to look at further. The density plot will not be made with the entire augmented subset. The rows of interest are the wrongly classified ones in this case. In figure 5.14 the distribution is shown. It doesn't show a complete gap between the hypotheses and premises for example. It is however noticeable that the right tail of the hypotheses curve stretches out a bit. It shows a lower density at the most dense part of the curve as well.

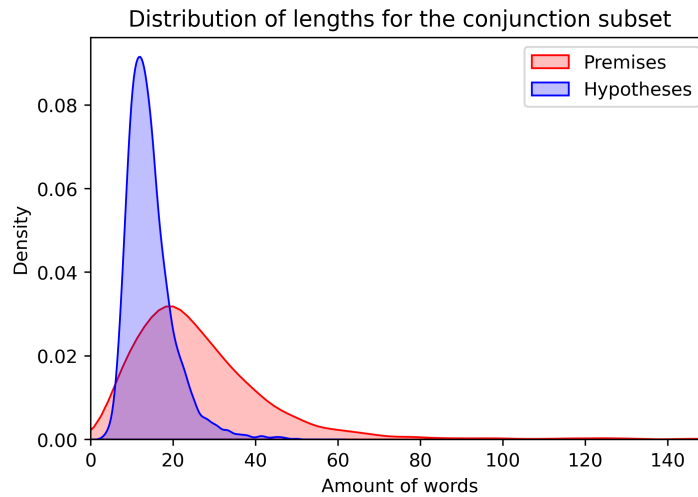


Figure 5.14: The distribution of the sentence lengths for the augmented conjunction subset. It is made with the rows that were wrongly classified

### 5.2.3 Augmented combination subset

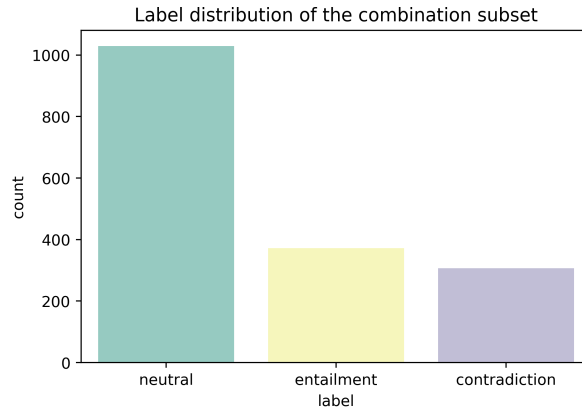
The augmented combination subset takes over some characteristics from the other two subsets. This is similar to what happened with the unedited subsets. There are some interesting rows that were wrongly classified. An example of this is given in table 5.15. The hypothesis is indeed changed. In this case the word that changed had impact on the change in label. This is because the whole premise is the piece of text in front of the comma in the hypothesis. Nothing changed before the comma so there difference is not made there. The word 'powerful' might have lots of different meanings but in this instance it would still qualify as an entailment pair.

Original	I live and die with them, that's right
Perturbed	I live and die with them, that's powerful
Premise	I live and die with them

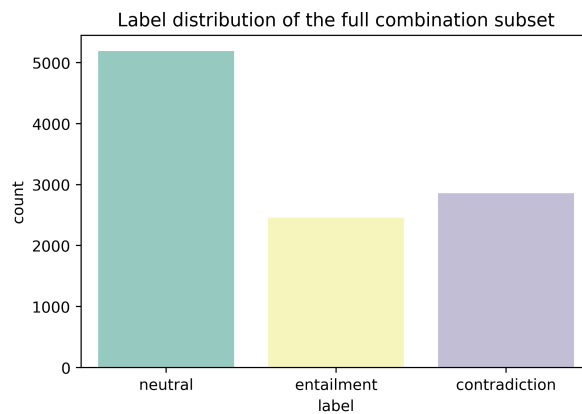
Figure 5.15: An example where the hypothesis and label changed. The sentence pair is thus wrongly classified here

This is a type of sentence pair that appears more often in this subset. The hypotheses are longer than the premises but these premises are short already. they are also partially or completely absorbed in the hypothesis. When a word outside of this part of the sentence is changed, the label changes to. However, the label shouldn't change.

The label distribution of the combination subset has a very high percentage of neutral labels being wrongly classified. The example in table 5.15 also falls in this category. Compared to all the subset rows together there are less contradiction labels than entailment. The contradiction labels seem to be wrongly classified less often than the other ones since the ratio shifts.



(a) Label distribution of the combination subset containing wrongly classified rows



(b) Label distribution of the full combination subset

Figure 5.16: Two label distribution plots with the upper one being the partial combination subset. The lower plot is the full combination subset

### 5.3 Comparison between unedited and augmented subsets

The unedited and augmented subsets have been reviewed so we can point out some differences and similarities. In the section there will be referencing back to figures and tables made for the results of each of the subsets. The first main difference that has to be mentioned is the difference in accuracy. In figure 5.17 the two types of subsets are plotted together to portray the difference in accuracy.

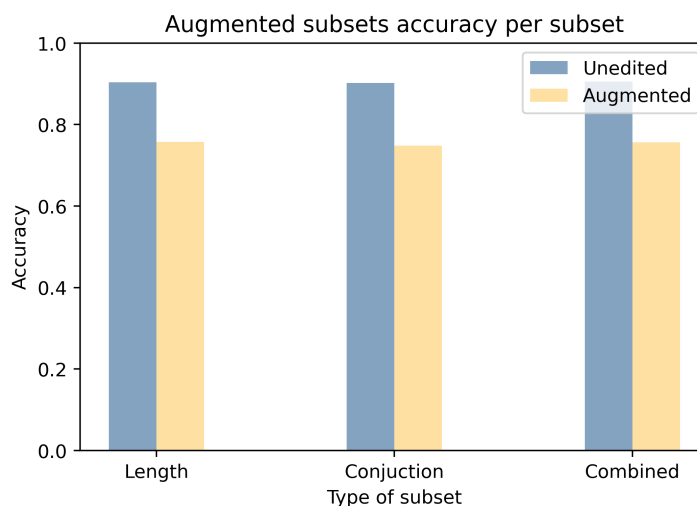


Figure 5.17: The accuracy of both the unedited and augmented subsets per subset

For both the unedited and augmented subsets, their inner differences are very small. There is no deviation of one particular subset dropping further in accuracy than one of the other subsets. We cannot say that one of the subsets significantly performs worse than another. This brings up the word 'worse'. It is established that accuracy drops across all subsets. The size of the drop is around 15% for each of the subsets. In the context of the subject this does seem to be quite a substantial drop. almost one fifth of the rows get wrongly classified. If we go back to the fake news articles that motivated this project it seems like a lot in many ways. Suppose that you see six fake news articles on the Facebook while scrolling for a while. Five of these six would be caught and blocked from your sight but there is one that remains. It could contain any type of mis-information as seen in figure 1.1.

A pattern of similarity that can be seen is the neutral label being wrongly classified the most. Even for the conjunction subset this is the case. It appears that it is most difficult for the model to deal with neutral labels. It is rather easily thrown off by a single word that changed which makes it change the label. Next to that, the distribution of sentence lengths stays approximately the same between the unedited and augmented subsets. For the augmented conjunction subset however it is visible that more hypotheses are longer. The density of at the most dense point is lower than it is for all the rows of the augmented conjunction subset.

## 6 Discussion

In the methodology and results there were assumptions and techniques used that might cause problems to the process. The assumption being made is the term 'worse' or in other words the vulnerability. It is stated that the accuracy drop is significant and the model is vulnerable to the augmented subsets. The explanation follows with an example of seeing fake news on a social media platform where one out of five or six are not seen as fake news. One might say that this is not that bad since the majority of the fake news articles is in fact picked up on and blocked. It seems to be impossible to detect and tackle every single piece of fake news. Even if the process is optimized through algorithms. In terms of percentage it is more difficult to say whether the model is vulnerable or not. Of course the aim is to see the highest possible drop in accuracy. Does that mean that only at a huge drop of for example 40% the model is vulnerable? The way to see these metrics always related back to the context of what is investigated. There will be other contexts where a 0.1% difference is a huge shift. Thus, it comes down to how many missed fake news articles and headers are too much? In our results, one out of five or six articles is deemed as a significant amount.

Furthermore, there is the choice of augmentation and quality control. There are many ways to augment a sentence having all sorts of operations. In our case, deletions and additions are disregarded because of the nature of the subsets. Next to that, the augmentation process also involves choosing a percentage of words to be swapped in the hypotheses. A higher or lower percentage would result in their own problems. A higher percentage would result in the meaning completely changing. This is unwanted as the goal is to try and break the NLP model with the minimum change. A lower percentage however would be so small that there wouldn't be any changes made to certain hypotheses. This leads to less augmented examples available to do the analysis on which is unwanted. The quality control of the augmented subsets is done with sentence similarity. The biggest assumption here is the threshold at 0.6 for the cosine similarity value. This is really a balance between letting more examples in with the risk of having wrong ones in there on one side. And on the other side the threshold could be more strict which possibly leads to less faulty examples but also less examples in general. Moreover, sentence similarity does not give a full quality guarantee either. It makes sure that the two sentences should overlap in meaning for some part but it does not fully exclude wrong sentences.

Related to this quality control is the fact a pitfall where labels change even though they should not. Imagine two sentences that have a high cosine similarity and pass the quality control. However, the augmented sentence does not have the same meaning anymore compared to the premise. This means that the original label is faulty. Any comparison to a predicted label would be worthless since the baseline label is wrong.

## 6.1 Future work

Of course, there is more follow-up research that can be done into the topic. One suggestion is to expand the amount of models that is being tested on. It is evident that DistilBERT is not the only NLP model out there and not the biggest nor best performing one overall. The reason why this model was chosen is also touched upon earlier in this paper due to its easier handling for the experiments. With bigger models such as it's bigger brother in a way, BERT, it would require more computing power to execute the technical part of the experiments. This is not the only model that could be used to see if it can be broken but it seems to be a logical follow-up step because the origin of the DistilBERT model.

Another remark that came up after and partially during the analysis as well is the way of augmenting datasets in order to try and break model. The augmentations done should be as minimal as possible but the augmented sentence should of course differ enough. For the experiments done in this paper an amount of rows was dropped after the sentence similarity method was applied. Sometimes also where it should not be necessary to do this. The ultimate goal after all is to make sure that the NLP model are pre-trained also on certain data that they are vulnerable to in order to eradicate this vulnerability. For this you need to have lots of data in this category and therefore it would be better not having to exclude data from a possible already smaller-than-usual amount.

Next to that, it might be interesting to try and see what happens when especially additions are used during the augmentation process as well. Here they are excluded but they could create an effect where longer hypotheses are increased even more. Another type of experiment that could be done is to try and create an artificial subset of hypotheses longer than their premises. This would mean that shorter hypotheses are turned into longer ones. There is the problem of changing the meaning of the hypothesis and the sentence pair altogether since something has to be added in this case. It would also most likely not just be one word but rather a part of a sentence. Based on how the subsets in this paper look that would indeed be the case.

## 7 Conclusion

The experiments done showed insight into the workings of the DistilBERT model regarding the classification of sentence-pairs. The augmentation done with TextAttack proved to be efficient at altering the subsets by replacing words in them. The quality was maintained by using sentence similarity. For this the cosine similarity was used. With already tough examples to go through, it performed quite well on both the unedited and augmented examples overall. Long hypotheses did not seem to have a baseline negative impact on the NLP model after all however. The fact that the data used is training data plays a part in the performance on the unedited subsets. The augmented subsets showed that the accuracy drops for each subset. Within the augmented subsets there is not a big difference in accuracy. This is the case for the unedited subsets too. The results of the comparison between the edited and unedited subsets showed that the accuracy dropped when the hypotheses were perturbed, also after the drop-out for possible wrong hypotheses.

We can look back at the main research question and think about how vulnerable they were. The accuracy drops does show vulnerability in the sense that we are missing out on possible fake news articles. The 15% decrease shows that we miss an amount of roughly one out of six fake news headers or articles. With social media platforms growing and growing this ratio seems to be rather significant. In the light of the context for this paper, we can answer the research question in the following way. The model DistilBERT is vulnerable to augmentations in datasets. There is no difference between the subsets in the amount of vulnerability they expose. The length, conjunction, and combination subset each show vulnerability. It has been shown that there are definitely ways to break a still novel NLP model by relatively simple augmentations to pieces of text. The field will continue to innovate and come up with new ways to break these models. Given everything that has been explored and answered, this paper can contribute to further research in the field.

## References

- Augmenter recipes commandline use.* (2020). Retrieved from [https://textattack.readthedocs.io/en/latest/3recipes/augmenter\\_recipes\\_cmd.html](https://textattack.readthedocs.io/en/latest/3recipes/augmenter_recipes_cmd.html)
- Belinkov, Y., Poliak, A., Shieber, S. M., Van Durme, B., & Rush, A. M. (2019). *On adversarial removal of hypothesis-only bias in natural language inference.* arXiv. Retrieved from <https://arxiv.org/abs/1907.04389> doi: 10.48550/ARXIV.1907.04389
- Bowman, S., Angeli, G., Potts, C., & Manning, C. (2016). *A large annotated corpus for learning natural language inference.* Stanford Linguistics, Stanford NLP Group and Stanford Computer Science.
- Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.* NAACL-HLT.
- Gleitman, L. R. (1965). *Coordinating conjunctions in english* (Vol. 41) (No. 2). Linguistic Society of America. Retrieved 2022-04-17, from <http://www.jstor.org/stable/411878>
- Glockner, M., Shwartz, V., & Goldberg, Y. (2018). *Breaking nli systems with sentences that require simple lexical inferences.* TU Darmstadt and Bar-Ilan University.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). *Explaining and harnessing adversarial examples.* arXiv. Retrieved from <https://arxiv.org/abs/1412.6572> doi: 10.48550/ARXIV.1412.6572
- Gururangan, S., Swamydipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. (2018). *Annotation artifacts in natural language inference data.* arXiv. Retrieved from <https://arxiv.org/abs/1803.02324> doi: 10.48550/ARXIV.1803.02324
- Kim, N., Patel, R., Poliak, A., Wang, A., Xia, P., McCoy, R., ... Pavlick, E. (2019). *Probing what different nlp tasks teach machines about function word comprehension.* John Hopkins University, Brown University, New York University, Google AI Language and Harvard University.
- Li, L., Shao, Y., Song, D., Qiu, X., & Huang, X. (2020). *Generating adversarial examples in chinese texts using sentence-pieces.* Fudan University.
- Minervini, P., & Riedel, S. (2018). Adversarially regularising neural NLI models to integrate logical background knowledge. *CoRR, abs/1808.08609*. Retrieved from <http://arxiv.org/abs/1808.08609>
- Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). *Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp.* arXiv. Retrieved from <https://arxiv.org/abs/2005.05909> doi: 10.48550/ARXIV.2005.05909
- Oshikawa, R., Qian, J., & Wang, W. Y. (2018). *A survey on natural language processing for fake news detection.* arXiv. Retrieved from <https://arxiv.org/abs/1811.00770> doi: 10.48550/ARXIV.1811.00770
- Saha, S., Nie, Y., & Bansal, M. (2020). *Conjnli: Natural language inference over conjunctive sentences.* arXiv. Retrieved from <https://arxiv.org/abs/2010.10418> doi: 10.48550/ARXIV.2010.10418
- Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter* (Vol. abs/1910.01108). Retrieved from <http://arxiv.org/abs/1910.01108>
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). *Intriguing properties of neural networks.* arXiv. Retrieved from <https://arxiv.org/abs/1312.4151>

- arxiv.org/abs/1312.6199 doi: 10.48550/ARXIV.1312.6199
- Wang, Z., Mi, H., & Ittycheriah, A. (2016). *Sentence similarity learning by lexical decomposition and composition* (Vol. abs/1602.07019). Retrieved from <http://arxiv.org/abs/1602.07019>
- Williams, A., Nangia, N., & Bowman, S. (2018). *A broad-coverage challenge corpus for sentence understanding through inference*. Association for Computational Linguistics. Retrieved from <http://aclweb.org/anthology/N18-1101>
- Yang, K.-C., Niven, T., & Kao, H.-Y. (2019). *Fake news detection as natural language inference*. arXiv. Retrieved from <https://arxiv.org/abs/1907.07347> doi: 10.48550/ARXIV.1907.07347
- Zhang, Y., Baldrige, J., & He, L. (2019). *Paws: Paraphrase adversaries from word scrambling*. arXiv. Retrieved from <https://arxiv.org/abs/1904.01130> doi: 10.48550/ARXIV.1904.01130
- Álvaro Figueira, & Oliveira, L. (2017). *The current state of fake news: challenges and opportunities* (Vol. 121). Retrieved from <https://www.sciencedirect.com/science/article/pii/S1877050917323086> (CENTERIS 2017 - International Conference on ENTERprise Information Systems / ProjMAN 2017 - International Conference on Project MANagement / HCist 2017 - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN/HCist 2017) doi: <https://doi.org/10.1016/j.procs.2017.11.106>



**Information to be provided by the student**

**i) Declaration concerning the TU/e Code of Scientific Conduct for the Bachelor's final project**

I have read the TU/e Code of Scientific Conduct.<sup>1</sup>

I hereby declare that my Bachelor's final project has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct.

Date

.....

Name

.....

Signature

.....



---

<sup>1</sup> See: <http://www.tue.nl/universiteit/over-de-universiteit/integriteit/wetenschappelijke-integriteit/>. The Netherlands Code of Conduct for Academic Practice of the VSNU can be found here also. More information about scientific integrity is published on the websites of TU/e and VSNU