Use of Natural language inference in optimizing reviews and providing insights to end consumers

Chahat Tandon
Computer Science and Engineering BIET
Davangere,India
chahat.7876@gmail.com

Pratiksha Jayesh Bongale
Computer Science and Engineering BIET
Davangere,India
pratikshajb@gmail.com

Arpita T M
Computer Science and Engineering BIET
Davangere,India
arpita.telkar@gmail.com

Sanjana R R
Computer Science and Engineering BIET
Davangere,India
sanjanarrdvg@gmail.com

Hemant Palivela
Digital Analytics, eClerx LLC,
Austin, Texas, US
hemant.datascience@gmail.com

Nirmala C R
Computer Science and Engineering BIET
Davangere,India
nirmala.cr@gmail.com

Article History: Submitted 7.1.2021 Revised 7.31.2021 Accepted 8.20.2021
Keywords: Natural Language Interface, Natural Language Processing, Deep Learning, Amazon Reviews

***Abstract*: Natural Language Interface or NLI has the potential to add syllogistic reasoning over the already existing facts and develop a new kind of knowledge dataset in itself. In this paper, we have demonstrated a Recognizing Textual Entailment wherein the task is to recognize whether a given hypothesis is true (Entailment), false (Contradiction) or unrelated(neutral) with respect to the sentence called premise. The task is performed by training MNLI corpus along with the manually collected dataset from Amazon Product Reviews each having hypothesis and premise pairs with corresponding labels. With this use case, we propose to bring sustainable development in the classification methods used by major E-commerce companies.**

I. Introduction

In the era of Artificial Intelligence, Deep Learning and Big Data having immense improvements, there is a need to expand Natural Language Processing (NLP) beyond what it does. Hence, expansion of NLP to perform different tasks requires different kinds of datasets and this leads to different types of challenges. One function among those which has lately gained need and popularity is the Natural Language Inference (NLI), also called Recognizing Textual Entailment (RTE). It is the task of defining whether the given hypothesis h is true, false or undetermined with respect to the given premise p. Hypothesis h is also considered as the conclusion c in many of the explanations. A fortunate NLI system is the one that exactly determines whether the hypothesis is entailed, contradicted or neutral to the given premise p. As per the discussion held by Condoravdi et al. (2003)[1] and others, a successful NLI is a suitable computation measure for a real natural language understanding. Goldberg and Hirst (2017)[2] and Nangia et al. (2017)[3] in their discussion clearly noted that solving NLI problems perfectly means to attain human level understanding of language. Hence, a continuous effort is put into designing a high level performing NLI model that has faster learning rate along with massive understanding capabilities.

Every product or service on the internet has a review system to know the opinion of their services from the customers and to help boost the customers loyalty towards their company. Study says that around 86% of the customers consider them as an indispensable resource when selecting the product. Deciding to purchase a product after going through its hundreds of reviews would be a time-consuming task as customers need to go through reviews of different products to find the best one of their choice. How do we reduce time consumption in this review analysis process without human resources? This paper aims at designing a system built using the Natural Language Inference (NLI), a branch of NLP, that will reduce the number of reviews you have to go through for a particular product to find out what needs to be known. The model helps in classifying the reviews into three easy categories namely, entailment, contradiction and neutral. By grouping the common subjects, it becomes convenient for the customer to identify the required review information; thus making it easier for them to make a decision.

There are diverse benchmarks designed for the numerous usecase of Natural Language Inference in different fields of business. A system which analyses the customer reviews is very much required for better functioning of the company and this paper is the first attempt on this use case to the best of our knowledge. The mode is built solely on the NLI aiming at reducing the number of reviews for the customer on the product/services (P/S) in amazon. This paper proposes a review analyzing system with the sole usage of NLI. Given a set of reviews in the paired format, we developed the model in such a way that it determines whether the pair of reviews stands true to each other, false or totally not relatable to one another. e.g *The material of the shirt is super soft. I just love it.* and *I'm inspired by the soft touch of this cloth.* are a pair of reviews on the same product which means the same. *Satisfied with the stitch of the shirt is clean and perfect.* and *Disappointed with the stitching done. Threads are coming out*. are another pair of reviews which clearly means opposite to one another. Identifying such pairs of reviews is necessary as we can group the common reviews keeping the rare one.

DOI reference number: 10.18293/JVLC2021-N1-011

II. Background

Natural Language Processing (NLP) consists of two sub-tasks namely Natural Language Understanding (NLU) and Natural Language Generation (NLG). Together they deal with the understanding of human's languages, processing them, analyzing them in an attempt to understand the semantics of the natural language sentences, and ultimately, generating an output back in human's language as was sent as input so that it is interpretable by humans effortlessly. NLP bundles a broad variety of use-cases, wherein some are considered to be easy tasks and the others a complex one to deal with that include recognition of entities [4] such as names of places, persons, etc., within texts, Sentiment Analysis [5], Machine Translation [6] of languages, Machine Reading Comprehension (MRC) [7], etc. Natural Language Processing requires an important task within itself, called Natural Language Inference (NLI). It is the task of appropriately inferring one sentence from another and classifying any two sentences (Premise and Hypothesis) in three categories namely, *entailment*, *contradiction*, and *neutral*. In some cases, some known facts and prior knowledge about the topic is taken into account while classifying the text pair. For instance, most Hindi speakers would know how to acknowledge a '*Namaste'* or *'Kya haal hai?'*, etc.

Multi-Genre Natural Language Inference (MNLI) [8], a larger corpus compared to SNLI roofs ten distinct genres of the English Language. It houses about 433k p-h (premise-hypothesis) pairs making it a good choice for NLI over English language. The test set has two categories namely, *matched set* - sentences with indistinguishable genres and *mismatched set* - sentences with distinguishable genres, thus, facilitating cross-genre language inference.

III. Related Work

Natural Language Inference is best dealt by applying Deep Learning based methods. We know for a fact that Deep Learning based methods require a humongous amount of training data to be able to learn the language representation and produce desired results. To acknowledge the massive need, many researchers, crowd workers, and others, came forward and created numerous datasets for NLI purposes. Hossein Amirkhani et al. [9] created the largest NLI dataset called FarsTail, entirely dedicated to the Persian language. The dataset consists of 10,367 examples that are given in both Persian language and an indexed-format to be beneficial for non-Persian experimenters. The Premise-Hypothesis pairs were created utilizing about 3500 multiple-choice queries with no or a little involvement of annotators with annotating the pairs. They investigated several techniques ranging from the traditional ones to the state-of-the-art methods bundling word embedding methods including word2vec [10], fastText [11], ELMo [12], BERT [13], and LASER [14], various modeling approaches specifically dedicated to Natural Language Inference dealing, such as, DecompAtt [15], ESIM [16], HBMP [17], ULMFiT [18], and cross-lingual transfer approaches. Another research conducted by Mohammad Mosharaf Hossain et al. [19] dealt with countering negation within English sentences as they are ubiquitous in common English sentences. Their study reveals that current datasets including the Stanford Natural Language Inference (SNLI) dataset, Recognizing Textual Entailment (RTE) dataset do not address negative words within sentences and which makes state-of-the-art transformers poor at handling words carrying negative meanings such as, no, nothing, never, not, isn't, haven't, hasn't, so on and so forth. The transformers tend to neglect the negative words or phrases and proceed with the inference task of classifying sentences into their respective classes.

In today's evolving world, code-mixing is prevalent. *Code-mixing* refers to the mixing of two or more languages while conserving with each other, for instance, *"Hey buddy! Haven't seen you since the first week of December! Kya chal raha hai?"*. This phrase consists of two languages namely, English and Hindi; this is how most people talk these days on social media platforms like WhatsApp, Instagram, Facebook, etc. Simran Khanuja et al. [20] went on and created a whole new dataset that takes into account the mixing of different languages. It consists of 400 code-mixed (English-Hindi) premises taken from 18 Bollywood movie scripts for which 2240 hypotheses are in the same format as premises, i.e., code-mixed.

Currently, there are many datasets that aid NLI with not just English but many more languages such as Hindi, Turkish, Spanish, German, Thai, Chinese, etc. There are some datasets that include premises and hypotheses just in English, for instance, SNLI; and some that include other languages such as French, Spanish, Greek, German, Swahili, Urdu, Arabic, etc., for instance, XNLI. The English-dedicated datasets include, *SICK (Sentences Involving Compositional Knowledge)* [21] which is one of the initial trials towards building larger datasets to support NLI functions. It bundles around 10,000 premise-hypothesis pairs in English language. The dataset was annotated for dual purposes, one, to determine correlation between sentences and two, entailment. The initial dataset consists of irregularly picked from about 8,000 ImageFlickr dataset along with the SemEval 2012 STS MSR-Video Description dataset. A few rule-grounded lexical and syntactic transformations have been put into every sequence of words to generate appropriate classification (entailment, contradiction, or neutral). Another well known corpus down the line is the *Stanford Natural Language Inference (SNLI)* [22] that tackles the need of huge annotated data for the NLI task to be solved using Deep Learning architectures. The corpus consists of around 5,70,000 labeled premise-hypothesis pairs out of which the training set consists of 550k samples, validation set of 10k, and test set of 10k examples. The entire corpus was collected via the Amazon Mechanical Turk. Every premise was asked to be paired with three different hypotheses, one entailment, one contradiction, and a neutral one. *MedNLI* [23], as the name suggests, is dedicated to the medicinal domain. An addition to the English dominated corpora is *SciTail*. SciTail [24] consists of science questions and answers as hypotheses and relevant word sequences from the web were taken as premises. A total of 1,834 queries taken together with about 16k neutral examples and around 10k entailment examples form the entire corpus. SciTail lacks the third label within NLI, the *contradiction* tag. *QA-NLI* [25], an automatically generated corpus built by leveraging the Question Answering datasets like the *SQuAD 2.0* [26]. The dataset followed the following pattern all over: correct answer - entailment, incorrect answer - contradiction, and unknown answer - neutral.

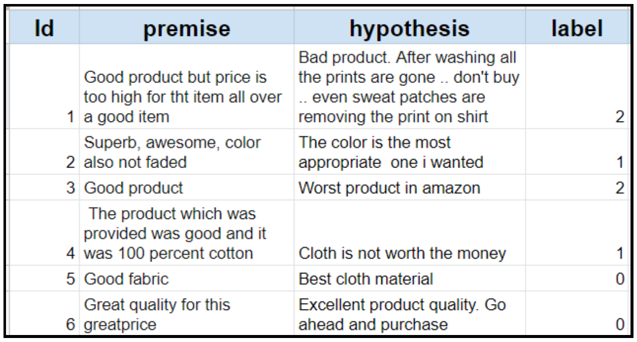
There are many Non-English corpora that have been designed to acknowledge NLI tasks in languages other than English including Evalita, ArbTEDS, German emails, etc. The Italian dataset, *Evalita* [27], consists of 800 short italian text pairs constituted using the Wikipedia articles. An Arabic language corpus, *ArbTEDS* [28], containing 600 annotated text pairs involving both inferable and non-inferable. Candidate duplets are taken from the web leveraging a nearly automatic instrument, with news headlines in Arabic as hypotheses and a passage rendered by Google's API for the just taken headline as the corresponding premise, annotated by eight annotators. *German emails* [29], a corpora built by emails sent by customers of a multimedia software company to its support hub as premises and the different class description as hypotheses. The matching classes relate to the entailment category (around 600) and non-matching classes map to the non-entailment category (around 21k). *ASSIN* [30], mixture of 10k couples, having both European Portuguese and Brazilian Portuguese pertaining to two different classes, namely, entailment and non-entailment. The massive *cross-lingual* dataset, XNLI which stands for *Cross-lingual Natural Language Inference*, the MultiNLI text pairs that were collected in a crowd-sourced manner, these pairs were then translated into 14 distinct languages by experts.

IV. Data Collection

A. *Dataset and Task to be performed:*

We handpicked 500 reviews, including one star to five-star reviews in order to give our data some uniformity. For instance, the review dataset for a particular shirt contained reviews ranging from its color, size, fabric quality to the fitting and the thread count. The dataset contained both positive as well as negative reviews. We then bifurcated the obtained reviews in three categories; entailment (0), neutral (1) and contradiction (2). Four columns were obtained containing the column id, the hypothesis, premise and the label. Consider the following review examples used for labelling the sentences:

For a book review: *"The book cover is beautiful!"* and *"Beauty lies in the cover as well as the content of the author."* can be labelled as **entailment (0)**.
For a phone review: "*I bought this phone for my father, and he liked it a lot, fast charge, good battery life*" and the "*screen width is small*" can be labelled as **neutral (1).**
For a laptop review: "*Bluetooth connection problem"* and "*Bluetooth connectivity is awesome but the battery drained fast*" can be labelled as a **contradiction (2)**.

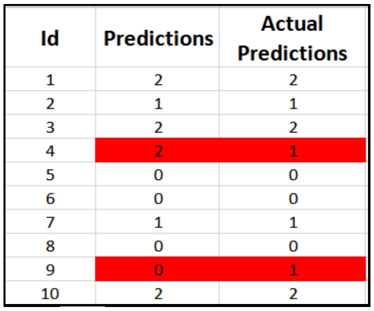


Id	premise	hypothesis	label
1	Good product but price is too high for tht item all over a good item	Bad product. After washing all the prints are gone .. don't buy .. even sweat patches are removing the print on shirt	2
2	Superb, awesome, color also not faded	The color is the most appropriate one i wanted	1
3	Good product	Worst product in amazon	2
4	The product which was provided was good and it was 100 percent cotton	Cloth is not worth the money	1
5	Good fabric	Best cloth material	0
6	Great quality for this greatprice	Excellent product quality. Go ahead and purchase	0

Fig. 1 - A snippet of training dataset

B. *Input and Output:*

Each dataset contains two sentences (a premise and a hypothesis) and a labeling class that indicates if the sentences describe the same thing (entailment), disagree with one another (contradiction) or talk about different things (neutral). So the model needs to take in two inputs (the two sentences) and return one of the three classes. For the output part, we feed in a submission xls file that saves the predictions made. A snippet of submission file has been shown below:



Id	Predictions	Actual Predictions
1	2	2
2	1	1
3	2	2
4	2	1
5	0	0
6	0	0
7	1	1
8	0	0
9	0	1
10	2	2

Fig 2- A snippet of Submission.xls

V. Method

A. *XLM-RoBERTa:*

Model used for this use case is XLM-RoBERTa. XLM-RoBERTa is based on RoBERTa which was proposed in 2019. This multilingual model is trained on filtered Common Crawl data across 100 different languages. The crawl data corpus is a collection of data in petabytes collected over 8 years of web crawling. In specific, we used xlm-roberta-large configuration of XML-RoBERTa. There are 24 hidden layers, 16 attention heads and 1024 hidden units.

We used TPUs in training. TPU short for *Tensor Processing Unit*, are application specific integrated circuits used to accelerate the workloads in machine learning. Being different from GPUs, a TPU needs to be initiated and set up to carry out work with the specified model in the notebook. Explicitly, a "strategy" needs to be demarcated regarding the working of the model and how it can be replicated across the eight GPU chips on the TPU board. The later part of these replica models being merged back together also needs to be taken care of once training has completed.

B. *Process Flow*

After setting the max length to 80; the batch size needed to be multiplied by the number of replicas (8).

This made sure that each of the eight GPU chips in the TPU was made to use the specified batch size and not one eighth of that number. The learning rate was set to 1e-5. The train and test set; each containing four columns namely, Id, hypothesis, premise and label was loaded. Augmentation of dataset helps increase diversity in the dataset all the while increasing accuracy. For this purpose, we used the Multi-Genre NLI Corpus dataset. The dataset was found to include 392702 rows × 3 columns. This proved to have added an advantage while classifying our dataset; thus improving the model's overall performance. The proposed system can be described by a process flow shown below:

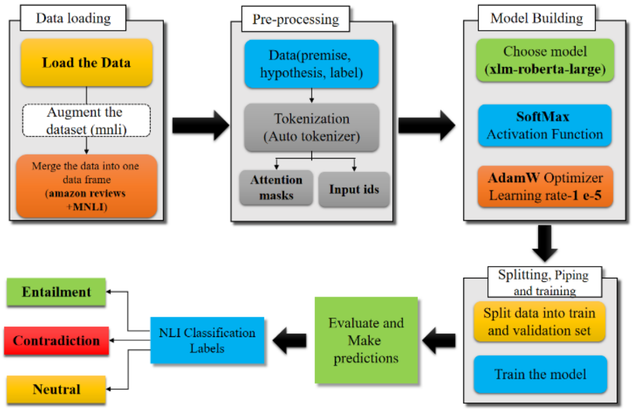

Data loading
Load the Data
Augment the dataset (mnli)
Merge the data into one data frame (amazon reviews +MNLI)
Pre-processing
Data(premise, hypothesis, label)
Tokenization (Auto tokenizer)
Attention masks
Input ids
Model Building
Choose model (xlm-roberta-large)
SoftMax Activation Function
AdamW Optimizer Learning rate-1 e-5
Splitting, Piping and training
Split data into train and validation set
Train the model
Evaluate and Make predictions
NLI Classification Labels
Entailment
Contradiction
Neutral

Fig 3- Process flow of the proposed model

C. *Training Phase*

In order to start with the training phase, we first **concatenated o**ur collected dataset with that of MNLI. In order for a machine to understand human-language, words need to be encoded and fed as integer inputs. This is known as **encoding**. In order to prepare sentences to be fed as input for training, the text is tokenized i.e. assigned numbers(tokens). Each model has its own unique set of tokens. Every sentence from the dataset is then converted from strings to arrays of tokens using Auto Tokenizer class from Transformers. The tokenizer even separates the words themselves into sub words. For instance, a word "bookmark" will be split as "book" and "mark" subwords. In our case, the premise and hypothesis both act as input and so the tokens from both will be merged into one array.

D. *Split and Pipeline*

The training set was split into 80% training and 20% validation set with a random state of 2020. Data Pipeline is used to exact every ounce of performance from the model at hand. We used Tensorflow data API to create a data pipeline in order to increase the performance while training. Commands like shuffling, slicing, prefetching the next batch, etc can be performed easily because of this pipeline. Lastly, the RoBERTa was added as the backbone of our model along with a softmax function layer in order to apply the correct class (entailment, neutral, contradiction or 0, 1, 2).

E. *Evaluations and Predictions*

The model is trained for 5 epochs with varying datasets each time and values for Training loss/ Accuracy as well as Validation loss/ Validation accuracy are noted. The new obtained predictions are saved in submissions.xls as output and can be later compared with the actual dataset. The below graphs showcase the evaluations obtained for the shirt dataset.

After the evaluation of individual datasets, a mixed dataset was created manually to check how the model would perform. It consisted partly of reviews of all the four products hand picked randomly. The model was deemed to have been trained properly as the validation accuracy obtained was 86%. Thus, we can conclude that the model seems to be ready to be used for **real-life business cases**.

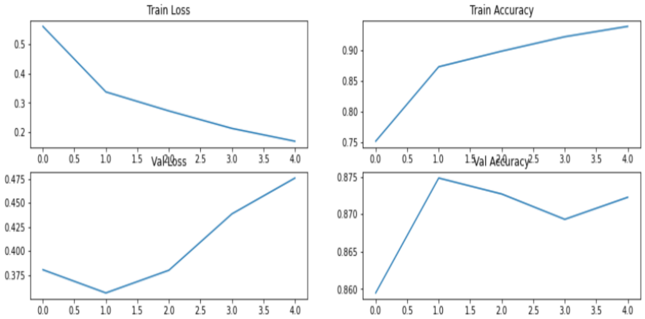


Fig 4- Evaluations obtained on shirt dataset.

VI. COMPARISION OF PREDICTIONS

The reviews collected for our dataset were of different products. 4 different products were chosen from amazon and all its reviews were collected. The first product we chose was the men's yellow shirt. Reviews for the same spoke about the material quality, buttons hook up, stitching done and various other features of the shirt. Regardless of whether the customer liked or disliked, all the reviews were selected to prepare our dataset. The NLI model minimized the number of similar reviews which would consume huge time in reading all of them. This minimization will not only help the customers in choosing their best product based on reviews easily and more sprightly, but will also help the sellers to look for the negative reviews for their product easily. This will assist the seller in improvising their shirt considering the negative reviews of the customer.

Similarly, the other products chosen were the crime and punishment book, Samsung M11 smartphone, Dell and HP laptops from amazon e-commerce website. All the possible reviews notwithstanding to positive and negative ones were collected for the same. Once the number of reviews were minimized by the NLI model by eliminating the reviews holding the same meaning, there remains fewer reviews making it easy for the customer to decide on purchasing their product quicker saving a huge amount of time for them. The creation of our model does not restrict to helping only the customers of the e-commerce website, it also makes things easier for the sellers online to understand customer needs and act faster and accordingly on it as the same in the above described shirt product. Any product chosen and review collected on the same, our NLI model designed for the reviews minimization eliminates one review from the one's, those are entailing. This system is found useful to all customers, sellers and manufacturers. Due to this magnificent feature of our model, this is considered to be a very good usecase in business. Such a type of model design is the first attempt in the field of NLP to the best of our knowledge.

A table showing Training/Validation and Testing accuracy on the aforementioned product review datasets has been showcased below. It is evident that the model was trained on a diverse variety of data and so was able to achieve great accuracy scores even when tested on different products. Thus this experiment can be deemed to be successful and useful for today's business cases.

Table 1– Training/Validation and Testing Accuracy

Reviews	Sets	Accuracy		
		Train	Validation	Test
Shirts	1-100	93.98	87.23	70
Book	101-200	94.18	87.39	60
Mobile	201-300	93.71	87.22	70
Laptop	301-400	94.04	87.1	79
	401-500	93.77	87.11	61

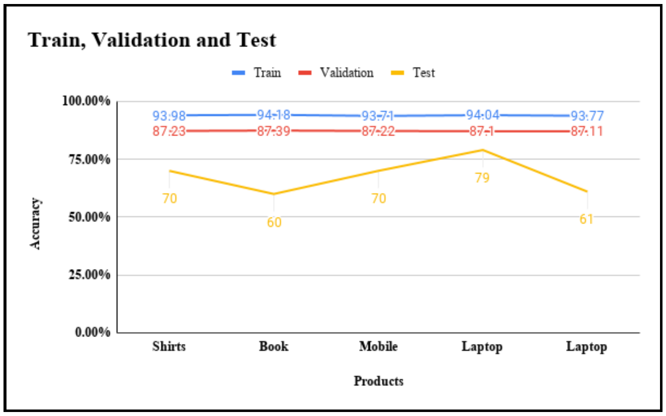


Fig 5 - Train, Validation and Test Accuracy on all the Products

VII. Conclusion

In this paper, we proposed the use of Natural Language Interface with respect to real life major E-commerce companies. We presented experiments on MNLI corpus along with manually obtained Amazon review dataset designed with 500 samples each having hypothesis and premise pairs with corresponding labels. The task of classifying hypotheses and premises in labels, i.e. true (Entailment), false (Contradiction) or unrelated (neutral) was done with 87% accuracy overall. The proposed model captures the need and possibilities with the use of NLI on a huge scale. This model caters to the need of classification strategy required in today's world of review driven sales. We hope that this model could be a stepping stone for future advancements in retail as well as general use cases.

References

[1] Cleo Condoravdi, Dick Crouch, Valeria De Paiva, Reinhard Stolle, and Daniel G Bobrow. 2003. Entailment, intensionality and text understanding. In Proceedings of the HLT-NAACL 2003 workshop on Text meaning-Volume 9, pages 38–45. Association for Computational Linguistics.

[2] Yoav Goldberg and Graeme Hirst. 2017. Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers.

[3] Nikita Nangia, Adina Williams, Angeliki Lazaridou, and Samuel Bowman. 2017. The RepEval 2017 Shared Task: Multi-Genre Natural Language Inference with Sentence Representations. In Proceedings of the 2nd Workshop on Evaluating Vector Space Representations for NLP, pages 1–10, Copenhagen, Denmark. Association for Computational Linguistics

[4] V. Yadav, S. Bethard, A survey on recent advances in named entity recognition from deep learning models, in: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 2145– 2158.

[5] A. Keramatfar, H. Amirkhani, Bibliometrics of sentiment analysis literature, Journal of Information Science 45 (1) (2019) 3–15.

[6] S. Yang, Y. Wang, X. Chu, A survey of deep learning techniques for neural machine translation (2020). arXiv:2002.07526.

[7] R. Baradaran, R. Ghiasi, H. Amirkhani, A survey on machine reading comprehension systems (2020). arXiv:2001.01582.

[8] A. Williams, N. Nangia, S. R. Bowman, A broad-coverage challenge corpus for sentence understanding through inference, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 1112–1122.

[9] Hossein Amirkhani, Mohammad Azari Jafari, Azadeh Amirak, Zohreh Pourjafari, Soroush Faridan Jahromi, Zeinab Kouhkan, FarsTail: A Persian Natural Language Inference Dataset, arXiv:2009.08820v1 [cs.CL] 18 Sep 2020.

[10] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Advances in neural information processing systems, 2013, pp. 3111–3119.

[11] P. Bojanowski, E. Grave, A. Joulin, T. Mikolov, Enriching word vectors with subword information, Transactions of the Association for Computational Linguistics 5 (2017) 135–146.

[12] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep contextualized word representations, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), 2018, pp. 2227–2237.

[13] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 4171–4186.

[14] M. Artetxe, H. Schwenk, Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond, Transactions of the Association for Computational Linguistics 7 (2019) 597–610.

[15] A. P. Parikh, O. Tackstrom, D. Das, J. Uszkoreit, A decomposable attention model for natural language inference, in: Proceedings of the 20 2016 Conference on Empirical Methods in Natural Language Processing, 2016, pp. 2249–2255.

[16] Q. Chen, X. Zhu, Z. Ling, S. Wei, H. Jiang, D. Inkpen, Enhanced LSTM for natural language inference, in: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017, pp. 1657–1668.

[17] A. Talman, A. Yli-Jyra, J. Tiedemann, Sentence embeddings in NLI with iterative refinement encoders, Natural Language Engineering 25 (4) (2019) 467–482.

[18] J. Howard, S. Ruder, Universal language model fine-tuning for text classification, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2018, pp. 328– 339.

[19] Mohammad Mosharaf Hossain, Venelin Kovatchev, Pranoy Dutta, Tiffany Kao, Elizabeth Wei, Eduardo Blanco, An Analysis of Natural Language Inference Benchmarks through the Lens of Negation.

[20] Simran Khanuja, Sandipan dandapat, Sunayana Sitaram, Monojit, Choudhury, A New Dataset for Natural Language Inference from Code-mixed Conversations, arXiv:2004.05051v2 [cs.CL] 13 Apr 2020.

[21] M. Marelli, L. Bentivogli, M. Baroni, R. Bernardi, S. Menini, R. Zamparelli, SemEval-2014 Task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment, in: Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland, 2014, pp. 1–8.

[22] S. R. Bowman, G. Angeli, C. Potts, C. D. Manning, A large annotated corpus for learning natural language inference, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 632–642.

[23] A. Romanov, C. Shivade, Lessons from natural language inference in the clinical domain, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 1586–1596.

[24] T. Khot, A. Sabharwal, P. Clark, SciTail: A textual entailment dataset from science question answering, in: Thirty-Second AAAI Conference on Artificial Intelligence, 2018, pp. 5189–5197.

[25] D. Demszky, K. Guu, P. Liang, Transforming question answering datasets into natural language inference datasets (2018). arXiv:1809. 02922.

[26] P. Rajpurkar, R. Jia, P. Liang, Know what you don't know: Unanswerable questions for SQuAD, in: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 784–789.

[27] J. Bos, F. M. Zanzotto, M. Pennacchiotti, Textual entailment at evalita 2009, Proceedings of EVALITA 2009 2 (6.4) (2009) 1–7.

[28] M. Alabbas, A dataset for Arabic textual entailment, in: Proceedings of the Student Research Workshop associated with RANLP 2013, 2013, pp. 7–13.

[29] K. Eichler, A. Gabryszak, G. Neumann, An analysis of textual inference in German customer emails, in: Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014), 2014, pp. 69–74.

[30] E. R. Fonseca, L. Borges dos Santos, M. Criscuolo, S. M. Alu´ısio, Overview of the evaluation of semantic similarity and textual inference, Linguamatica 8 (2) (2016) 3–13.

[31] A. Conneau, G. Lample, R. Rinott, A. Williams, S. R. Bowman, H. Schwenk, V. Stoyanov, XNLI: Evaluating cross-lingual sentence representations, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, 2018, pp. 2475–2485.