Deep Learning for Stance Detection: A Review and Comparison of the State-of-the-Art Approaches

Jacob Roeters van Lennep

27 - June - 2021

Abstract

Stance detection is a Natural Language Processing task that can detect if the input text is in favour, against or neutral towards a target. Research on stance detection has been growing and evolving over the last decade. In this paper, the current approaches for stance detection are discussed with a focus on the deep learning approaches. The organized competitions are discussed, and the most used traditional and deep learning approaches are shown. The challenges that arise with deep learning approaches are looked into further. Finally, an experiment was performed to examine and demonstrate the effects of a small data set on various stance detection models, this was done using the SVM, CNN, and BERT models on the SemEval 2016 data set. This experiment shows that a smaller data set has a greater negative impact on the CNN model than the SVM model. BERT is affected the least and outperforms the other models significantly.

1 Introduction

Over the recent years, stance detection has grown into one of the most prominent subfields within the field of Natural Language Processing (NLP). This surge in popularity is evident when one looks at the number of publications. While in the period of 2006–2010 only five articles were published on this topic, this skyrocketed to 38 and 78 in the periods of 2015–2016 and 2017–2019, respectively [25].

One reason for the increase in popularity of stance detection is its relation with Fake News detection. Over the last years, the term 'Fake News' has entered the public vocabulary, and we have witnessed its impact on the public, political and medical domains. Some estimates show that Fake News inflicts a global yearly cost of 78 billion.[1] However, companies have not created a reliable way of assessing Fake News and handling it properly. A building block for fake news detection is stance detection. Therefore it is logical that the increasing attention on fake news detection also increases the interest in stance detection.

At the beginning of 2020, a notable survey by Küçük and Can [25] showed the state of stance detection up until 2019. To complement this, we will analyze and discuss the state of thee field of the last two years. In the last two years the field has shifted towards more deep learning approaches, but to give a complete overview, both the traditional and the deep learning approaches that are currently used are examined.

This paper is organized as follows. Section 2 definitions that are used throughout the paper are introduced. The formulated research question is shown in Section 3. In Section 4

[1] https://www.zdnet.com/article/online-fake-news-costing-us-78-billion-globally-each-year/

some of the most important competitions in the field of stance detection are discussed and the increasing popularity of deep learning approaches in stance detection is demonstrated. Additionally, a comparison will be drawn between traditional and deep learning approaches, with a specific focus on performance. The methods used for stance detection are discussed in Section 5. In Section 6 the challenges for the deep learning approaches are highlighted. The experiment to look into the challenge of data set size for deep learning is shown in Section 7. The results of the experiment are shown in Section 8. The paper is concluded with the discussion about the experiment in Section 9, Section 10 where the conclusion and recommendation of future work are given and the final Section, Section 11 responsible research.

2 Background

In the past years, a lot of papers were published about stance detection. However, there is no unanimity in definition and naming across articles. Therefore, it is important to clearly outline the definitions that will be utilised in this paper.

2.1 Defining Stance Detection

Stance detection consists of three main areas: generic stance detection, rumour stance classification, and Fake News stance detection. This paper will focus predominantly on generic stance detection. There are two common ways of approaching the generic stance detection, Multi-Target stance detection and Cross-Target stance detection. All definitions are provided below.

Generic Stance Detection *Generic stance detection is the most common form of stance detection. It can be seen as an classification task. The input is a piece of text and a target. The stance detection than classifies the piece text in three main categories how it relates to the target. The most common categories are Favor, Against, Neutral. With these categories the stance of the text is decided.*

Rumour Stance classification *In rumour stance classification the input is a piece of text and a rumour pair. This form looks it whether the topic of the text is inline with the rumour or not. The two main categories are Supporting and Denying. Some also add Querying and Commenting to extend the two main categories.*

Fake News Stance Detection *This form of stance detection is a form where the headline of an news item and the actual text of a, not necessarily the same, news item are used as input. The Fake News stance detection classifies the relation between the headline and the text of the article in four main categories, Agrees, Disagrees, Discusses, Unrelated. With Discusses meaning that both text and headline are about the same topic.*

Multi-Target Stance Detection "*For an input in the form of a piece of text and a set of related targets, multi-target stance detection is a classification problem where the stance of the text author is sought as a category label from this set: {Favor, Against, Neither} for each target and each stance classification (for each target) might have an effect on the classification for the remaining targets*" [25, page 2]. This paper will use this definition as we discuss multi-target stance detection. All future stance detection references would be to multi-target stance detection.

Cross-Target Stance Detection "*Cross-target stance detection is a classification problem where the stance of the text author is sought for a specific target as a category label from this set: {Favor, Against, Neither}, in a settings where stance annotations are available for (though related but) different targets, i.e. there is not enough stance-annotated training data for the target under consideration*" [25, page 2]. When referring to this definition, cross-target stance detection is explicitly mentioned.

3 Research Questions

In this paper, we examine the current field of stance detection by an extensive literature research and by conducting an experiment in which the influence of data set size is compared between traditional and deep learning approaches. The following research question and sub-questions have been formulated.

Main question:

- What is the state-of-the-art on deep learning approaches for stance detection and how do these deep learning approaches compare to the feature-based approaches?

Sub-questions:

1. What deep learning approaches are currently used in Stance detection ?
2. How do the architectures of commonly used stance detection models that use deep learning approaches look like?
3. Are deep learning approaching gaining popularity in the field of Stance Detection?
4. How do deep learning approaches compare in performance to traditional stance detection approaches?
5. What is needed to accelerate the use of deep learning approaches in stance detection?

4 Competitions

In 2016 two major stance detection competitions, SemEval task six [34] and NLPCC-ICCPOL task four [57], sparked the increase of attention towards stance detection. Shortly after, in 2017, another competition, IberEval [41], was organized. One of the most recent competitions on stance detection is the SardiStance challenge organized during the Evalita in 2020 [5]. In the following section, the competitions are described and the approaches used are mentioned. In table 1 an overview of the deep learning approaches versus the traditional approaches by the submitted teams is given.

4.1 SemEval 2016

Task six of the 2016 edition of SemEval was solving a stance detection problem. Mohammad et al. [34] explain the outcome of task six. The task consists of two subtasks, focusing on stance detection on a data set consisting of tweets. Task A consists of five targets, Atheism (AT), Climate change is a real concern (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalization of Abortion (LA). Task B has only one target, Donald Trump (DT). This data set is further described in Section 7.2.

Subtask A is a traditional supervised classification task with a given annotated data set. 19 teams submitted a model for this task. In this subtask, the baseline model, which uses an SVM approach, achieved a F_{avg} score [34, Section 4.3] of 68.98, outperforming all the submissions. The winner of this subtask is MITRE [59], which uses two RNN classifiers, one based on LSTM units. The MITRE team achieves a F_{avg} of 67.82.

Subtask B is a weakly supervised framework. In this task, there was no provided training data. The data consist of tweets related to Donald Trump. Subtask B received nine submissions. The organisation provided an SVM based baseline that achieved a F_{avg} of 28.43. In this subtask, the models of 8 of the teams outperformed the baseline. The winning team pkudblab [54] achieved a F_{avg} score of 56.28. They used a CNN based on the architecture discussed by Kim [23].

In this competition, the participants mainly used traditional approaches. Out of the 16 participants who published their models, only four used deep learning approaches. The team MITRE of Zarella and Marsh [59] was the only team that used an LSTM. This team was outperforming all other teams in the competition. Three other teams, pkudblab [54], Tohoku [20], and DeepStance [52] all used a CNN approach. These approaches had a F_{avg} of 67.33, 62.21, and 63.54, respectively, on subtask A.

4.2 NLPCC-ICCPOL 2016

In the NLPCC-ICCPOL 2016 conference, task four is dedicated to stance detection. Xu et al. [57] describe the two subtasks of the competition. In this competition, the data set was consisting out of text from Weibo, a Chinese microblog website. Task A was a supervised task where 16 teams participated. Task A highest score was 0.7106, which team RUC_MMC achieved. Their approach was a combination of an SVM and a random forest. Task B was an unsupervised task where five teams participated. The highest performing team was team March, which achieved a F_{avg} of 0.4687.

In this competition, out of the 16 teams that compete, four published about their models. From these four publications, only one used a deep learning approach. The team of Yu et al. [58] used a technique that is based on a LSTM model. With this approach, they scored a F_{avg} of 0.5656 on task A.

4.3 IberEval 2017

In the IberEval conference, a combined competition of stance and gender detection was held [48]. In this competition, a stance regarding the independence of Catalonia needed to be detected based on Spanish and Catalan tweets. Ten teams participated in this competition. The best performing team on the stance subtask was iTACOS [27]. This team combined multiple non-deep learning techniques. Their approach consisted of an SVM, Random Fores, Logistic Regression, Decision Tree and a multinomial Naive Bayes. iTACOS achieved a F_{score} of 0.4901 on the Spanish tweets and a F_{score} of 0.4885 on the Catalan tweets.

Out of the ten participating teams, nine of them published their models. Out of these nine teams, four used some form of deep learning in their models. For example, the team deepyCybErNet [41] created three different approaches based on RNN, LSTM, and GRU. DeepyCybErNets GRU approach scored a F_{score} of 0.3066 on the Spanish, which was the highest of their three approaches. On the Catalan tweets, the LSTM approach achieved the highest score with a F_{score} of 0.379.

The team ELiRF-UPV [15] combined an LSTM and CNN approach and achieved a F_{score} of 0.5510 on the Spanish tweets, earning a fourth place on the stance detection task.

Table 1: Deep learning aproaches versus traditional approaches in competitions

	Deep learning approaches	Traditional approaches	Total
SemEval 2016	4	**12**	16
NLPCC-ICCPOL 2016	4	**12**	16
IberEval 2017	4	**5**	9
EVALITA 2020	**10**	2	12

The team of Wojatzki and Zesch [55], LTL_UNI_DUE had a non deep learning classifier (SVM) and a deep learning classifier (LSTM). On both the Spanish and Catalan data set, the SVM outperformed the LSTM approach. The last team, from Ambrosini and Nicolo [2] used an LSTM approach, achieving a F_{score} of 0.556.

4.4 EVALITA 2020

In the EVALITA 2020, the competition SardiStance was held. [5] This competition also consisted of two subtasks. Task A was a textual stance detection, and task B is a contextual stance detection task. In this competition, twelve teams participated. Out of these twelve teams, ten used deep learning approaches. In task A the team UNITOR [14] achieved a F_{avg} of 0.6853. They achieved this result by implementing a model based on BERT. On task B, IXA was the best performing scoring a F_{avg} of 0.7445.

5 Stance Detection Approaches

There are three main categories of approaches in stance detection; Traditional machine learning approaches, deep learning approaches, and ensemble learning approaches.

The traditional machine learning approaches have been, in a historical view, the most popular. However, currently, the deep learning approaches are the most popular. Therefore in the coming section, an overview of the essential traditional machine learning approaches (Section 5.1) and the deep learning approaches (Section 5.2) are given.

For more insight on ensemble learning approaches the article by Küçük and Can gives a good overview of the most popular approaches in this category [25, Section 5.3].

5.1 Traditional approaches

In the 2016 SemEval challenge described in Section 4.1 traditional approaches were evidently the most popular category of approaches. Traditional approaches are often also feature-based machine learning approaches. In these approaches most commonly used feature extraction are bag-of-words, ngrams, skip-gram, hash-tags. [21] Also word vector representations are common, popular ones are word2vec [33] and GloVe vectors [37].

Support Vector Machine

The most popular traditional approach is the support vector machine (SVM). This approach is used in most approaches published in 2016. SVM is also used as the baseline approach of the SemEval 2016 challenge, which scored the highest F_{score}. The winning team of the NLPCC-ICCPOL also used an SVM based approach.

For the SemEval 2016 baseline, the organizers used three different SVM approaches for the baseline. An SVM-unigrams, which consisted of five SVM classifiers. One per target, see Section 7.2 for an explanation of the targets. These classifiers used word unigram features.

The other two SVM that were used for the SemEval Baseline were SVM-ngrams and SVM-ngrams-comb. SVM-ngrams also consisted out of five SVM classifiers using n-grams and character n-grams features. The SVM-ngrams-comb consisted of only one SVM trained over all the five targets. The SVM-ngrams had the highest F_{avg} of 68.98 [34].

Another SVM that was used in the SemEval 2016 challenge was the one from Patra et al. [36]. Their approach to creating the features for the SVM was based on a bag-of-words. They made different individual topic bags that were related to one of the targets. Besides the topic bags, they had two different kinds of lexicons. One that already existed and one they created themselves. Patra et al. used the Stanford Parser3 to elicit dependency relations between words to improve the features. They created five different SVMs, one per target, and achieved a F_{avg} score of 60.60.

Logistic Regression

Logistic regression is the second most popular approach from the traditional approaches. For example, logistic regression is used in Zhang and Lan [61] which they achieved a F_{avg} score of 65.55. Their feature method was a combination of topic, similarity, sentiment lexicon, tweet specific, and word vector feature.

Others

Besides the SVM and Logistic Regression, other traditional approaches are also seen as the bases for stance detection. Naive Bayes and decision trees are being the most widely used after the SMV and LR. In Appendix A Table 6 shows 12 other approaches and related articles.

5.2 Deep learning approaches

In the two competitions in 2016, the number of deep learning approaches was relatively small: around 25% of the approaches used deep learning. However, the deep learning approaches gained popularity in the following years. In 2018, with the arrival of BERT [7], the division shifted. In the stance detection challenge of 2020, described in Section 4.4, most of the approaches were deep learning approaches, and most of these were using a form of BERT. Other deep learning approaches that are still popular are LSTM, RNN, and CNN.

BERT

The Bidirectional Encoder Representations from Transformers (BERT) model was first introduced by Devlin et al. in 2018 [7]. It outperformed most common deep learning approaches in different NLP tasks. The BERT model has two stages. The first being the *pre-trained stage* where the BERT model is trained on unlabeled data for different NLP tasks. The second stage, the *fine-tuning stage* will be task-specific. It will, in this stage, train on labelled data related to the specific task.

BERT is used in several recent papers on stance detection. In the Evalita 2020 task A, team Unitor [14] ranked first with the use of a BERT model. They based their model of BERT on a RoBERTa architecture proposed by Liu et al. [31]. The model that Liu et al.

submitted is an adoption of UmBERTo that is pre-trained on a subset of the OSCAR corpus [35] to use it on Italian tweets specifically.

Wani et al. [53] use BERT only to create the mathematical vector they use in different deep learning approaches. The two approaches that achieve the highest F_{avg} on the SemEval data set are *bert-cnn*, a score of 68.4 and *bert-nn* with a score of 67.17.

Other BERT approaches can be found in Conforti et al. [6], Popat et al. [38], Samih and Darwish [43], Grimminger and Klinger [16], and Alkhalifa and Zubiaga [1].

In more elaborated tasks beyond stance detection, BERT is also used in combination with other deep learning approaches. Kula et al. [26] combine a BERT model with an RNN approach to detect Fake News. Kaliyar et al. [22] combine a BERT model with a CNN approach.

LSTM

Before introducing BERT, an approach that was the most popular was the long-short term memory (LSTM) network [19]. This network is based on an RNN architecture. LSTMs in stance detection are often built out of two separate LSTM models, one for seeing if the input is related to the topic and one for, when it is associated with the topic, to see if it is in favour or against it.

LSTMs are used in Augenstines et al. [3], Dey et al. [8], Zarrella and Marsh [59], Vinayakumar et al. [41], Gonzalez et al. [15], Wojatzki and Zesch [55], Rajendran et al. [40], and Alkhalifa and Zubiaga [1].

CNN

Convolution Neural Network (CNN) is an approach that is also used often in stance detection. The basis for most implementations of CNNs in stance detection is based on architecture Kim [23] proposed in 2014 for sentence classification tasks. In the SemEval challenge, the pkudblab [54] team used a CNN implementation, achieving a second place on task A and first place on task B.

Other usage of CNNs are described in Igarashi et al. [20], Vijayaraghavan et al. [52], Gonzales et al. [15], Ambrosini and Nicolo [2], Zhang et al. [60], and Zhou et al. [62].

Others

Besides BERT, LSTM, and CNN, different deep learning approaches are used in stance detection. Recurrent Neural Network (RNN) is used not only as an LSTM but also with different variants. RNNs are used in approaches described in Vinayakumar et al [41] and Benton and Dredze [4]. Another architecture that is used is Gated recurrent units (GRU) this approach is used in Hiray and Duppada [18], Wei et al. [54], and Zhou et al. [62]. A recent new approach is based on a Graph Convolutional Network (GCN), this approach is described by Liang et al. [29].

6 Challenges for Deep Learning Approaches

Two main problems arise with the use of deep learning approaches in stance detection.

The first problem is a lack of large-scale annotated data sets that can be used for stance detection. This can be a problem since deep learning approaches need a rather large data set to achieve better performance than traditional machine learning approaches.

The second problem that the deep learning approaches introduce is explainability. With traditional machine learning approaches, it is more transparent how a model comes to a particular outcome. However, with deep learning approaches, this is not that easy. Therefore, when stance detection is used as a building block for fake news detection, then the explainability of the model could be an essential aspect. As moderating decisions will be made with these models, the models need to be explained and not solely be black boxes.

7 Experiment

As stated in the previous section, the size of a data set may impact the performance of deep learning algorithms. To further explore this impact, and how this may compare to traditional approaches, an experiment is conducted. Three models are used to conduct this experiment. On all three models, SVM, CNN, and BERT, three different data sets will be used to train these models. The first data set being the full SemEval 2016 data set, the second and third being 75% and 50% of the full data set. After this training, F_{scores} are compared from a test set; this set will stay the same on all three occasions.

The models will be described in Section 7.1 and altered data sets that are based on the SemEval 2016 data set. These data sets are describe in Section 7.2. The results of this experiment are presented in Section 8.

7.1 Models

We can observe the effect on the two different approaches by using three models, two deep learning models (CNN and BERT) and a traditional model (SVM), and changing the data set size. SVM has been the most popular model of the traditional approaches. On the other hand, BERT is the most used and highest performing approach in most stance detection models currently used.

SVM

The SVM model used during this experiment is created by Vass[2]. This is a basic SVM model with specific pre-processing to achieve the highest results. The different pre-processing techniques that have been tried can be found in the paper by Vass[51]. For this experiment, the SVM is used over a pre-processed data set that used TF-IDF on unigrams.

CNN

The CNN model that is is used for the experiment is based on the CNN architecture proposed by Kim [23]. It has been altered to specific preform on the SemEval data set by the highest performing team in the SemEval challenge, pkudblab [54]. The code has been made available by the team.[3]

This model is further improved by Ghosh et al. [12]. In their comparative study, they created a new method of pre-processing and used hyperparameter tuning in a five-fold cross-validation on the training set. The final hyperparameters are shown in Table 2. Ghosh et al. used the same voting scheme as the pkudblab team used. The experiment will also use this voting scheme and hyperparameters in this experiment.

[2]https://gitlab.ewi.tudelft.nl/cse3000/2020-2021/rp-group-65/rp-group-65-kvass.git
[3]https://github.com/wan-wei/SE16-Task6-Stance-Detection

Table 2: Hyperparameters of the CNN model

Hyperparameter	**Value**
Dropout	0.5
Learning_rate_decay	0.95
Squared norm limit (AT, LA, FM)	7
Squared norm limit (CC, HC)	8

Table 3: Hyperparameters of the BERT model

Hyperparameter	**Value**
Learning_Rate	2e-5
Num_Train_Epochs	50
Warmup_Proportion	0.1
Max_Seq_Length	128

BERT

The BERT model that will be used in the experiment is based on the paper by Devlin et al. [7]. In the comparison paper of Ghosh et al. [12] a BERT model is implemented. This is the pre-trained model that will be used in this experiment. The model can be found on their GitHub repository[4]. The BERT model is a pre-trained uncased large 24-layered model, which can be found on GitHub.[5] During the experiment, this model has been run in Google Colab with a TPU[6] set up. It is configured with the hyperparameters shown in table 3.

7.2 Data Set

The SemEval 2016 data set [34] consists of 2900 labelled tweets as training data, having a stance towards five targets. The targets are Atheism (AT), Climate change is a real concern (CC), Feminist Movement (FM), Hillary Clinton (HC), and Legalization of Abortion (LA). It also has a test data set for all five targets. The data set can be found online.[7]

Every data record consists of a tweet, a target, and a stance label (Favor, Against, None). However, in this training data set, not all three stance labels are equally represented.

Table 4 shows the distribution of tweets between the five targets. Data set two and three are taking a percentage of the data records based on the stance label to preserve the same division of stance labels in the training set while shrinking it.

The first, data set one, is the complete data set of the SemEval 2016 task six challenge. The second, data set two, is a reduced set. All the stance label have randomly reduced to 75 % of the original set. The third, data set three, has been further reduced. Therefore, data set three only contains 50% of the records of data set one.

[4]https://github.com/prajwal1210/Stance-Detection-in-Web-and-Social-Media
[5]https://github.com/shalmolighosh/bert/
[6]https://cloud.google.com/tpu/docs/quickstart
[7]https://www.saifmohammad.com/WebPages/Stancedataset.htm

Table 4: Distribution of stance labels in data sets for the experiment

	Data set 1					Data set 2					Data set 3				
	AT	CC	FM	HC	LA	AT	CC	FM	HC	LA	AT	CC	FM	HC	LA
Favor	92	212	210	112	105	69	159	158	84	79	46	106	105	56	53
Against	304	15	328	361	334	228	12	246	271	251	152	8	164	181	167
None	117	168	126	166	164	88	126	95	125	123	59	84	63	83	82
Total	513	395	664	639	603	385	297	499	480	453	257	198	332	320	302

8 Results

Table 5 shows the results of the experiment. The scores shown are F_{scores}. The highest F_{score} per data set, per target, is in bold. The displayed numbers are the F_{score} also used and described by Mohammad et al. [34]. All three models are compared on the same three data sets. Overall, the BERT model outperforms almost all other models on every target. The two exceptions are that the CNN model outperforms the BERT model on Hillary Clinton (HC) on data set one and data set two.

The results show two clear and exciting patterns. First, the performance score of BERT on smaller data sets is comparable to that of larger data sets. It is generally assumed that deep learning approaches have a worse performance on smaller data sets. This data contradicts this assumption, as BERT, deep-learning-based approach, outperforms the SVM and CNN model even for the smaller data sets.

Second, when comparing the SVM and CNN model, a clear difference can be seen between two data sets. On data set one, the CNN model outperformance the SVM on every target. On the experiment of data set two, this is not the case anymore. On the targets Atheism (AT), Climate Change (CC), the SVM model scores higher than the CNN model. The presented results on data set three also show that the SVM performs better on two of the targets. This could be because these two targets have the least of labeled data in the sets cmpared to the other targets.

9 Discussion

As stance detection is being used in online content moderation it is important to see how this field changes. One of the main problems has been the lack of bigger data sets that are annotated. Therefor, it is essential to see how the deep learning approaches, that have been used more, perform on smaller data sets. This section reflects how the experiment is conducted and what could be improved. The experiment conducted in this paper has been carried with the utmost care. However, there are still some methodological factors that may potentially improve the power and validity of this experiment. The three main focus points are the results, the experiment setup and the data set that is used. A more in-depth look at the results that are mentioned in Section 8 is given and discussed. The setup of the experiment and the used data set is also discussed.

9.1 Results

When critically looking at the results of the experiment, the F_{scores} are very close to each other. This makes the impact of the smaller data set less clear and not necessarily significant.

Table 5: Results of the experiment

	Data set 1					
	AT	CC	FM	HC	LA	F_{avg}
SVM	0.5699	0.3834	0.5196	0.5931	0.5860	0.5304
CNN	0.6281	0.4453	0.5409	**0.6780**	0.6657	0.5916
BERT	**0.7778**	**0.8452**	**0.5964**	0.5903	**0.6786**	**0.6977**
	Data set 2					
	AT	CC	FM	HC	LA	F_{avg}
SVM	0.5905	0.4567	0.5224	0.5951	0.6094	0.5548
CNN	0.5419	0.4466	0.5279	**0.6027**	0.6675	0.5573
BERT	**0.7778**	**0.8393**	**0.6179**	0.5903	**0.6786**	**0.7008**
	Data set 3					
	AT	CC	FM	HC	LA	F_{avg}
SVM	0.5968	0.3878	0.5104	0.5375	0.6165	0.5298
CNN	0.5009	0.4363	0.5486	0.4365	0.6173	0.5079
BERT	**0.7870**	**0.7798**	**0.6107**	**0.6875**	**0.7071**	**0.7144**

This could be improved to do an cross fold validation with the training and testing set. This would lead to more stable results and perhaps also a wider difference between the results of the CNN model and the SVM model.

9.2 Experiment setup

In this experiment, there are three models. These models are chosen for two reasons. The first was the performance in stance detection competitions, where they performed well—the second being online availability. There are, however, other approaches that would be interesting to compare. For example, the LSTM architecture would be another deep learning approach that could be added to the experiment. For a traditional approach, Linear regression or naive Bayes could be considered.

Adding extra models could give more insight into how the traditional approach compares to the deep learning approaches. This will improve the robustness of the experiment.

The BERT model used in the experiment is based on an older BERT architecture[8] this is the BERT model from 2019. This is done because there were already hyperparameters published by Ghosh et al. [12] which made the use of this model and hyperparameters more accessible. However, the new BERT model [50] would likely outperform the older one.

9.3 Data set

The data set that was used in the experiment is one of the most popular data sets of stance detection. A lot of challenges and research has been using this data set. Therefore, it was a logical choice to use this well know data set to conduct this experiment.

To improve the experiment, a more extensive data set could be used. This will increase the difference in data volume between the first, second, and third sets and will most likely

[8]https://github.com/shalmolighosh/bert/

result in a clearer difference between the deep learning and traditional approaches when changing the size.

Another possibility to get a more concrete result is doing this experiment over multiple data sets. This would increase the reliability.

10 Conclusions and Future Works

The field of stance detection has shifted from traditional approaches to deep learning approaches. To further explore the current focus of the stance detection field, we examined the most important contests and challenges that have been held in the past, giving insight into the most popular and current approaches for stance detection. Next, an overview was given of the most popular approaches. In essence, our contribution shows the state-of-the-art on stance detection.

To look at one of the common hurdles in deep learning approaches, an experiment is conducted. This experiment looks at the impact of data size between traditional approaches and deep learning approaches. When the data set size is halved, the SVM model performs similar to the CNN model. Another clear conclusion from the experiment is that the BERT model had comparable performance when analyzing normal and downsized data sets, contradicting the assumption that deep learning approaches are unsuited for smaller data sets.

BERT is currently leading in stance detection, and any future research would be most likely to include the BERT model. As new smaller BERT models were published after the 2019 variant used in this experiment, it would be interesting to see how these models perform compared to the old BERT model. Another research area could be the impact of different data sets. As the used data set is relatively small, the impact of larger data sets could be investigated.

11 Responsible Research

There are two main focusess of responsible research. The first being the reproducibility and biases. Both will be discussed in this section.

11.1 Reproducibility

The experiment performed in this paper consists of three models and a data set. The three models used, the SVM, CNN, and BERT model, are all available online. Together with the hyperparameters mentioned in Section 7.1 these models can be run exactly as the models were used in this paper. In Section 7.1 all the repositories where the models can be found and downloaded are presented.

The data set that is used in the experiment, the SemEval 2016 task 6 data set, is also available online. In Section 7.2 the link to this data set can be found. A random sample can be taken from the full data set to create the second and third data sets. As these samples are random there is likely a change that the results from a random sampled set will differ from the one used in this experiment. Therefore this has a negative effect on reporducibility and is a thread to the validity.

By having all the fundamental parts for the experiment readily available and a clear description of how the experiment is performed. The only problem in the reproducibility is the random sample for data set two and data set three.

11.2 Biases

As stance detection approaches are based on machine learning, biases are an essential aspect that needs to be addressed. As stance detection could be used in Fake News detection, the models must be biased free as there could be wrongful results when there are biases in the stance detection models. These erroneous results could impact society, as news that should be labelled as false will be labelled as true and vice versa.

To prevent a biased model, the data set needs to examine. A problem could occur if the data set is skewed towards a specific stance being opposed or in favour. In table 4 the distribution of the data set that is used in this experiment is shown. Some targets are not as uniformed distributed as one would like.

For the sake of the experiment conducted in this paper, the bias implications are not severe. As the research focuses on seeing how models F_{scores} compare to one another, the social impact of the biases within the data set is slim. To be able to use these models in other applications more repeated experiments with different data sets should be performed.

Acknowledgement

This project would not be possible if Pradeep Murukannaiah was not willing to be our supervisor and responsible professor. Thank you for the guidance, feedback and support. I would also like to thank the group I have been working with during the setting up of this project and the writing and performing the experiment. This group consisted of Wout Haakman, Kristof Vass, Abel van Steenweghen and Simon Marien.

A Other traditional approaches

Table 6: Other traditional approaches

Approach	**Article(s)**
ILP	Ghosh et al. [13], Konjengbam et al. [24], Li et al. [28].
kNN	Shenoy et al. [46].
log-linear model	Ebrahimi et al. [10].
maximum entropy	Hercig et al. [17], Xu et al. [56].
FastText	Rohit and Singh [42].
Stochastic Gradient Descent	Lozhnikov et al. [32].
k-means clustering	Simaki et al. [47].
matrix factorization	Lin et al. [30], Qiu et al. [39], Sasaki et al. [44].
factorization machines	Sasaki et al. [45].
Multiple Convolution Kernel Learning	Tsakalidis et al. [49].
statistical relational learning	Ebrahimi et al. [11].
weakly-guided learning	Dong et al. [9].

References

[1] Rabab Alkhalifa and Arkaitz Zubiaga. Qmul-sds @ sardistance: Leveraging network interactions to boost performance on stance detection using knowledge graphs, 2020.

[2] L. Ambrosini and Giancarlo Nicolò. Comparative study of neural models for the coset shared task at ibereval 2017. In *IberEval@SEPLN*, 2017.

[3] Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. Stance detection with bidirectional conditional encoding, 2016.

[4] Adrian Benton and Mark Dredze. Using author embeddings to improve tweet stance classification. In *Proceedings of the 2018 EMNLP Workshop W-NUT: The 4th Workshop on Noisy User-generated Text*, pages 184–194, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[5] Alessandra Cignarella, Mirko Lai, Cristina Bosco, Viviana Patti, and Paolo Rosso. Sardistance @ evalita2020: Overview of the task on stance detection in italian tweets. In *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12 2020.

[6] Costanza Conforti, Jakob Berndt, Marco Basaldella, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. Adversarial training for news stance detection: Leveraging signals from a multi-genre corpus. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*, pages 1–7, Online, April 2021. Association for Computational Linguistics.

[7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[8] Kuntal Dey, Ritvik Shrivastava, and Saroj Kaushik. Topical stance detection for twitter: A two-phase lstm model using attention. In Gabriella Pasi, Benjamin Piwowarski, Leif Azzopardi, and Allan Hanbury, editors, *Advances in Information Retrieval*, pages 529–536, Cham, 2018. Springer International Publishing.

[9] Rui Dong, Yizhou Sun, Lu Wang, Yupeng Gu, and Yuan Zhong. Weakly-guided user stance prediction via joint modeling of content and social interaction. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 1249–1258, 2017.

[10] Javid Ebrahimi, Dejing Dou, and Daniel Lowd. A joint sentiment-target-stance model for stance classification in tweets. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 2656–2665, 2016.

[11] Javid Ebrahimi, Dejing Dou, and Daniel Lowd. Weakly supervised tweet stance classification by relational bootstrapping. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1012–1017, 2016.

[12] Shalmoli Ghosh, Prajwal Singhania, Siddharth Singh, Koustav Rudra, and Saptarshi Ghosh. Stance detection in web and social media: A comparative study. *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, page 75–87, 2019.

[13] Subrata Ghosh, Konjengbam Anand, Sailaja Rajanala, A Bharath Reddy, and Manish Singh. Unsupervised stance classification in online debates. In *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, pages 30–36, 2018.

[14] Simone Giorgioni, Marcello Politi, Samir Salman, R. Basili, and Danilo Croce. Unitor @ sardistance2020: Combining transformer-based architectures and transfer learning for robust stance detection. In *EVALITA*, 2020.

[15] José González Barba, Ferran Pla, and Lluís Hurtado Oliver. Elirf-upv at ibereval 2017: Stance and gender detection in tweets. In *IberEval@ SEPLN*, pages 193–198, 01 2017.

[16] Lara Grimminger and Roman Klinger. Hate towards the political opponent: A twitter corpus study of the 2020 us elections on the basis of offensive speech and stance detection, 2021.

[17] Tomás Hercig, Peter Krejzl, Barbora Hourová, Josef Steinberger, and Ladislav Lenc. Detecting stance in czech news commentaries. In *ITAT*, pages 176–180, 2017.

[18] Sushant Hiray and Venkatesh Duppada. Agree to disagree: Improving disagreement detection with dual grus, 2017.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[20] Yuki Igarashi, Hiroya Komatsu, Sosuke Kobayashi, Naoaki Okazaki, and Kentaro Inui. Tohoku at SemEval-2016 task 6: Feature-based model versus convolutional neural network for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 401–407, San Diego, California, June 2016. Association for Computational Linguistics.

[21] Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Stanford University, 2020.

[22] Rohit Kumar Kaliyar, Anurag Goswami, and Pratik Narang. Fakebert: Fake news detection in social media with a bert-based deep learning approach. *Multimedia Tools and Applications*, 80(8):11765–11788, Mar 2021.

[23] Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.

[24] Anand Konjengbam, Subrata Ghosh, Nagendra Kumar, and Manish Singh. Debate stance classification using word embeddings. In *International conference on big data analytics and knowledge discovery*, pages 382–395. Springer, 2018.

[25] Dilek Küçük and Fazli Can. Stance detection: A survey. *ACM Comput. Surv.*, 53(1), February 2020.

[26] Sebastian Kula, Michał Choraś, and Rafał Kozik. Application of the bert-based architecture in fake news detection. In Álvaro Herrero, Carlos Cambra, Daniel Urda, Javier Sedano, Héctor Quintián, and Emilio Corchado, editors, *13th International Conference*

on Computational Intelligence in Security for Information Systems (CISIS 2020), pages 239–249, Cham, 2021. Springer International Publishing.

[27] Mirko Lai, Alessandra Cignarella, and Delia Hernandez Farias. itacos at ibereval2017: Detecting stance in catalan and spanish tweets. In *IberEval 2017*, volume 1881, pages 185–192, 09 2017.

[28] Chang Li, Aldo Porco, and Dan Goldwasser. Structured representation learning for online debate stance prediction. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3728–3739, 2018.

[29] Bin Liang, Yonghao Fu, Lin Gui, Min Yang, Jiachen Du, Yulan He, and Ruifeng Xu. Target-adaptive graph for cross-target stance detection. In *Proceedings of the Web Conference 2021*, page 3453–3464, April 2021.

[30] Junjie Lin, Wenji Mao, and Yuhao Zhang. An enhanced topic modeling approach to multiple stance identification. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pages 2167–2170, 2017.

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[32] Nikita Lozhnikov, Leon Derczynski, and Manuel Mazzara. Stance prediction for russian: data and analysis. In *International Conference in Software Engineering for Defence Applications*, pages 176–186. Springer, 2018.

[33] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[34] Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41, San Diego, California, June 2016. Association for Computational Linguistics.

[35] Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online, July 2020. Association for Computational Linguistics.

[36] Braja Gopal Patra, Dipankar Das, and Sivaji Bandyopadhyay. JU_NLP at SemEval-2016 task 6: Detecting stance in tweets using support vector machines. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 440–444, San Diego, California, June 2016. Association for Computational Linguistics.

[37] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[38] Kashyap Popat, Subhabrata Mukherjee, Andrew Yates, and Gerhard Weikum. STANCY: Stance classification based on consistency cues. In *Proceedings of the 2019*

Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6413–6418, Hong Kong, China, November 2019. Association for Computational Linguistics.

[39] Minghui Qiu, Yanchuan Sim, Noah A Smith, and Jing Jiang. Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *Proceedings of the 2015 SIAM international conference on data mining*, pages 855–863. SIAM, 2015.

[40] Gayathri Rajendran, Bhadrachalam Chitturi, and Prabaharan Poornachandran. Stance-in-depth deep neural approach to stance classification. *Procedia Computer Science*, 132:1646–1653, 2018. International Conference on Computational Intelligence and Data Science.

[41] Vinayakumar Ravi, Sachin Kumar S, Premjith B., Prabaharan Poornachandran, and Soman Kp. Deep stance and gender detection in tweets on catalan independence@ibereval 2017. In *IberEval@ SEPLN*, pages 222–229, 09 2017.

[42] Sakala Venkata Krishna Rohit and Navjyoti Singh. Analysis of speeches in indian parliamentary debates. *arXiv preprint arXiv:1808.06834*, 2018.

[43] Younes Samih and Kareem Darwish. A few topical tweets are enough for effective user stance detection. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2637–2646, Online, April 2021. Association for Computational Linguistics.

[44] Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. Other topics you may also agree or disagree: Modeling inter-topic preferences using tweets and matrix factorization. *arXiv preprint arXiv:1704.07986*, 2017.

[45] Akira Sasaki, Kazuaki Hanawa, Naoaki Okazaki, and Kentaro Inui. Predicting stances from social media posts using factorization machines. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3381–3390, 2018.

[46] Gourav G Shenoy, Erika H Dsouza, and Sandra Kübler. Performing stance detection on twitter data using computational linguistics techniques. *arXiv preprint arXiv:1703.02019*, 2017.

[47] Vasiliki Simaki, Carita Paradis, and Andreas Kerren. Stance classification in texts from blogs on the 2016 british referendum. In *International Conference on Speech and Computer*, pages 700–709. Springer, 2017.

[48] M. Taulé, M. A. Martí, F. M. R. Pardo, P. Rosso, C. Bosco, and V. Patti. Overview of the task on stance and gender detection in tweets on catalan independence. In *IberEval@SEPLN*, 2017.

[49] Adam Tsakalidis, Nikolaos Aletras, Alexandra I Cristea, and Maria Liakata. Nowcasting the stance of social media users in a sudden vote: The case of the greek referendum. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 367–376, 2018.

[50] Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Well-read students learn better: On the importance of pre-training compact models, 2019.

[51] K.K Vass. Preparing stance detection: Feature extraction methods and their performance used for feature-based machine learning algorithms. unpublished, 2021.

[52] Prashanth Vijayaraghavan, Ivan Sysoev, Soroush Vosoughi, and Deb Roy. DeepStance at SemEval-2016 task 6: Detecting stance in tweets using character and word-level CNNs. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 413–419, San Diego, California, June 2016. Association for Computational Linguistics.

[53] Mudasir Ahmad Wani, Nancy Agarwal, and Patrick Bours. Impact of unreliable content on social media users during covid-19 and stance detection system. *Electronics*, 10(1), 2021.

[54] Wan Wei, Xiao Zhang, Xuqin Liu, Wei Chen, and Tengjiao Wang. pkudblab at SemEval-2016 task 6 : A specific convolutional neural network system for effective stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 384–388, San Diego, California, June 2016. Association for Computational Linguistics.

[55] Michael Wojatzki and Torsten Zesch. Neural, non-neural and hybrid stance detection in tweets on catalan independence. In *IberEval@ SEPLN*, pages 178–184, 09 2017.

[56] Kang Xu, Sheng Bi, and Guilin Qi. Semi-supervised stance-topic model for stance classification on social media. In *Joint International Semantic Technology Conference*, pages 199–214. Springer, 2017.

[57] Ruifeng Xu, Yu Zhou, Dongyin Wu, Lin Gui, Jiachen Du, and Yun Xue. Overview of nlpcc shared task 4: Stance detection in chinese microblogs. In *Natural Language Understanding and Intelligent Applications*, volume 10102, pages 907–916, 12 2016.

[58] Nan Yu, Da Pan, Meishan Zhang, and Guohong Fu. Stance detection in chinese microblogs with neural networks. In *Natural Language Understanding and Intelligent Applications*, volume 10102, pages 893–900, 12 2016.

[59] Guido Zarrella and Amy Marsh. MITRE at SemEval-2016 task 6: Transfer learning for stance detection. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 458–463, San Diego, California, June 2016. Association for Computational Linguistics.

[60] Shaodian Zhang, Lin Qiu, Frank Chen, Weinan Zhang, Yong Yu, and Noémie Elhadad. We make choices we think are going to save us: Debate and stance identification for online breast cancer cam discussions. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 1073–1081, Republic and Canton of Geneva, CHE, 2017. International World Wide Web Conferences Steering Committee.

[61] Zhihua Zhang and Man Lan. ECNU at SemEval 2016 task 6: Relevant or not? supportive or not? a two-step learning system for automatic detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 451–457, San Diego, California, June 2016. Association for Computational Linguistics.

[62] Yiwei Zhou, Alexandra Cristea, and Lei Shi. Connecting targets to tweets: Semantic attention-based model for target-specific stance detection. In *International Conference on Web Information Systems Engineering*, pages 18–32, 10 2017.