Building task-specific stance identification models: an evaluation of the active learning approach

Lisa Vasileva
Edoardo Guerriero
e2.vasileva@student.vu.nl
e.guerriero@student.vu.nl
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

Isa Maks
Kasper Welbers
isa.maks@vu.nl
k.welbers@vu.nl
Network Institute, Vrije Universiteit Amsterdam
Amsterdam, The Netherlands

ABSTRACT

Training models to perform stance identification requires large annotated corpora that are not always available, especially in a research setting where researchers must collect and annotate data from scratch. Active learning is a technique for reducing the number of texts that are required to train a model, but while it has been studied for decades, most of the papers that test the effectiveness of this approach in a natural language processing scenario rely on already made and well known datasets like the IMDB Movie Reviews Dataset [9]. In this paper we test different active learning approaches in a scenario where task-specific training data needs to be developed from scratch. We collected and annotated tweets about vaccination to build our own dataset, and trained a model using different active learning pool-based strategies to see if active learning can be a viable strategy for building task specific stance identification models.

1 INTRODUCTION

Sentiment analysis, opinion mining and stance identification are natural language processing tasks that aim at identifying, extracting and labeling affective and subjective information from unstructured text data [8]. In the last years deep learning has been widely used in sentiment analysis [25] and deep models quickly established themselves as the state of the art for many sentiment analysis objectives.

One downside of machine learning algorithms is the need for a big amount of annotated documents, or labelled data, in order to achieve a good level of accuracy and generalization. This is especially true for recent deep learning architectures which are quickly replacing classic shallow models in almost every Natural Language Processing (NLP) task. Relying on human annotators to create sufficient labeled data is not always feasible, because document annotation is an expensive, time-consuming task.

In this paper we investigate *active learning* (AL) as a viable strategy for reducing the workload of creating sufficient labelled data for training a sentiment classifier. AL is a subfield of machine learning that studies different ways for selecting which documents need to be labeled first in order to improve the learning rate of the model, and thereby reduce the number of labeled documents that is required to train a good model. We explore whether established AL strategies are also effective for the complex task of sentiment classification in cases where only a moderate number of labeled documents is available. If this is the case, then AL strategies could be a powerful addition to the toolkit of researchers that require task specific models for their research, though additional research would have to determine whether non-random selection of a small number of training cases does not bias the model. If the boons of AL are small or neglect-able, then we would argue that a true random selection of training cases has preference.

To test the effectiveness of AL strategies in our use case, we collected and annotated a new corpus, and use this corpus to apply the AL framework with different pool-based sampling strategies. We have resorted to tweets as textual data because Twitter is an active and popular platform for expressing opinions. Regarding the content of the tweets, we focus on the vaccination topic, a debate-rich and socially volatile topic, that is likely to garner a plethora of opinions, especially on a social platform like Twitter.

The task we choose to perform is stance detection. The motivation behind this focus is stance detection being valuable and beneficial for various applications: information retrieval, text summarization, and textual entailment [10]. We define stance as stated in SemEval competition: "Stance detection is the task of automatically determining from text whether the author of the text is in favor of, against, or neutral towards a proposition or target."[10]. We will better stress out the peculiarities of stance detection compared to the generic definition of sentiment analysis in the next sections.

The sampling strategies we decided to compare are: uncertainty sampling (focusing on the Shannon entropy), cost-effective active learning [22], deep Bayesian active learning [23] [18] and a random sampling baseline. The choice relies on the possibility to easily implement those strategies without extensive experience in machine learning and hence making them available to researchers across a range of disciplines (e.g., Humanities and Social Sciences).

2 RELATED WORKS

2.1 Machine learning for Sentiment analysis and Stance Detection

Early works on automatic stance detection usually compared rule-based and classic machine learning classifiers like SVM or Naive Bayes [20][19] [21]. These works show that it is possible indeed to train a model to perform stance detection, but even using complex hand crafted features, the best accuracy obtained hardly overcome the threshold of 75% (using domain specific datasets), with unigram baselines achieving 65%. This proves that stance detection is a task still far from being solved.

More recent works tried to apply deep learning models. In particular, the work of Kim [6] tested different types of convolutional neural networks with several datasets for sentiment analysis, i.e. classification of sentences as positive-negative or neutral, showing that CNN can obtain results closer to optimized traditional algorithms. Moreover, CNN can achieve such results using only pre-trained embedding vectors (e.g. word2vec), without the need for complex syntactic or semantic features extraction.

During the 2016 SemEval workshop a shared task focused on stance detection in tweets was proposed [10]. An interesting outcome of the workshop is that none of the participants were able to outperform the baseline consisting in a Support vector machine trained with unigrams and bigrams. It is worth to stress out though that lot of teams tried to apply deep learning architectures. And this trend of increasing use of deep architecture in publications is true in all NLP branches [24]. Since the scope of this paper is to test the available resources and the usefulness of active learning in a real research scenario, and not to try to overcome state of the art results, we decided to use a classic architecture as the multi-channel CNN proposed in [6] and a Naive Bayes classifier as a comparison baseline. A detailed description of the CNN architecture is provided in section 3.

Algorithm 1: Active Learning general framework

Input :

(1) Training dataset (labeled documents)
(2) Set of unlabeled documents for active selection
(3) Sampling strategy
(4) N, number of documents to select after each training
(5) θ, accuracy threshold to reach

while *Model accuracy* $\geq \theta$ **do**

(1) Train a model using the current training dataset
(2) Apply model to unlabelled data (to get probabilities score)
(3) (optional) Calibrate predictions
(4) Apply sampling strategy to select N documents to annotate
(5) Ask human annotator to annotate previously selected documents
(6) Expand training dataset including previously annotated documents

end

2.2 Active Learning

In the statistical literature, AL was first introduced as an optimal experimental design problem [3]. The key idea behind AL is the possibility to design better strategies than random sampling (sometimes called *passive learning*) for the selection of training instances subsequently used to train a learning algorithm. These strategies are designed to select training examples which are hard for a model to predict or classify. By retraining a model with the addition of few relevant training instances, it is possible to achieve greater accuracy than by adding more training instances selected randomly.

The pseudo code in Algorithm 1 shows how to expand the typical training framework of a supervised machine learning model to include AL. We can see that the main elements required are:

(1) A set of labeled documents: required to train an initial model
(2) A set of unlabeled documents: required to sample from it new training instances
(3) A sampling strategy: required to rank the unlabelled documents in term of the confidence of the model predictions

In this work, we focused on *uncertainty sampling* strategies, introduced in [7]. The choice derived from the nature of the problem we tried to tackle, which is a classification task, and from the choice of the model, which is a probabilistic one.

These strategies rank a set of unlabelled data by leveraging the predicted probabilities of belonging to a specific class generated by model we're interested to retrain. These probabilities can be either combined in a unique score by applying some function to them, or they can be directly used to rank the data. The *entropy* uncertainty sampling (just Entropy from now on) apply the Shannon entropy function [17] to the probabilities of the unlabelled data. The resulting entropy score is then used to rank in descending order the data (instances with high entropy implies also a higher uncertainty of the model on those instances). Other classical uncertainty sampling strategies instead look only at the raw probabilities. The *least-confident* selection rank the unlabelled data by picking for each instance the highest probability, while instead the *max-margin* selection compute for each unlabelled data the difference between the highest and second highest probability. A brief example of the differences between the above mentioned strategies is shown in table 1.

Scores	Tweet 1	Tweet 2	Tweet 3
P(favor)	.70	.69	.68
P(against)	.20	.15	.16
P(neutral)	.10	.16	.16
Entropy	.80	.83	**.84**
Least-conf	.70	.69	**.68**
Max-margin	**.50**	.53	.52

Table 1: Example of classic uncertainty sampling selection. In bold the scores associated to the tweets that would be selected by the strategy.

Other famous active learning strategies other than uncertainty sampling include *query-by-committee* [16], which are more theoretically motivated, but they also require to train a set of several models (a committee), making the choice of these active learning framework less practical. For a complete survey of classic active learning sampling strategies for shallow models please refer to [15].

While the above mentioned strategies have been deeply tested with shallow machine learning algorithms, their effectiveness when applied to deep architectures is still not clear. This is probably due to the fact that deep models are less calibrated than shallow models as proved in [5], i.e. despite the fact that the models can achieve

higher accuracy, the probability scores they provide are less reliable, usually overestimated. It is worth to mention that in the same paper the authors prove that it is possible to improve the calibration of the probability scores of a deep model by applying post-processing techniques like temperature scaling *temperature-scaling*. Since this technique do not impose any special requirement in terms of models architecture, we decided to include it in our experimental setting.

Despite the issues that deep models bring with them, there have already been several attempts to successfully combine them with AL. For example, an interesting new active learning framework called *Cost-Effective-Active-Learning* (CEAL), has been recently proposed and tested with CNN in a image recognition task [22]. In CEAL, classic uncertainty sampling strategies are used to select the most relevant examples from a set of unlabelled instances. The novelty lies in adding to the classic AL framework an extra step, which is to select not only elements with a low level of prediction confidence, but also elements with a high level of confidence. The latter will not be presented to a human annotator, instead, they will be labelled automatically using the model prediction. Interestingly enough, Wang et al. do not take into consideration the calibration problem in their paper, but their results prove the effectiveness of the new framework, at least in their experimental setting.

The last sampling strategy we interested to try is *Bayesian active learning by disagreement*. It has been first proposed for an image classification task in [4], and then applied to several NLP tasks in [18]. From the latter paper we took also the abbreviation DO-BALD, which stands for Monte-Carlo Bayesian active learning by disagreement. This strategy takes advantage of the regularization technique called *dropout*, usually implemented in deep models to avoid overfitting. Dropout is typically performed only during model training. For each layer of weights to learn, a specific amount of them are randomly selected during each backward step and temporary 'dropped', i.e. they are not updated for that specific training iteration step. During the test phase, all the weights are kept active and halved. The basic idea of Bayesian active learning is to keep the dropout active also during test phase. In this way, it is possible to perform a Monte-Carlo sampling of several predictions for the same instance we are interested to classify, since the random selection of weights to use to predict the instance will produce different prediction at each sampling. The set of predictions generate for a specific instance can then be used to estimate the parameters of their underling distribution, like their mean and variance. In an active learning setting, the variance of the distributions calculated for each unlabelled instance can be used as a measure of the uncertainty of the model, and therefore used to rank the data to select the ones with higher variance.

3 METHODOLOGY

3.1 Data collection

To investigate the use of active learning for building task-specific stance classification models, we collected and annotated a novel corpus of tweets on the topic of vaccination. In addition to annotations of the stance of a tweet, we looked into linguistic, pragmatic and topical peculiarities; the analysis thereof can be found in section 3.3 below.

Stance	Annotator			
	1	2	3	4
FAVOR	144	137	115	151
AGAINST	191	213	‘212	184
NEUTRAL	65	50	73	65

Table 2: Stance distribution per annotator

Only fact	Annotator			
	1	2	3	4
yes	42	74	103	85
no	358	326	297	315

Table 3: Fact/non-fact distrubution per annotator

Confidence	Annotator			
	1	2	3	4
0	14	19	3	1
1	66	9	10	7
2	78	20	29	34
3	79	96	95	78
4	71	199	65	195
5	92	57	198	85

Table 4: Confidence scores per annotator

The total number of tweets is 1019, which were manually annotated by the authors of the paper. 400 of all tweets were annotated jointly in order to calculate inter-annotator agreement for the annotation effort as well as inspect points of uncertainly in the data, and the test of the tweets were annotated separately by two authors according to the annotation guidelines specifically established for the project.

3.2 Data collection strategies

To collect the data we first resorted to the strategy described in SemEval competition. This strategy included polling the Twitter API with query hashtags, intended to gather tweets of specific stance: favor, against and stance-ambiguous hashtags (for each of six following topics: 'Atheism', 'Climate Change is a Real Concern', 'Feminist Movement', 'Hillary Clinton', 'Legalization of Abortion', 'Donald Trump'). These tweets were then manually annotated for stance with crowdsourcing platform, and the resulting annotations were inspected for inter-annotator agreement. For a tweet to be included in the final dataset, 60% of crowd annotators had to agree on stance. [11]

In our case, adopting the strategy from SemEval-2016 and collecting data with query hashtags split into favour, against and stance-ambiguous hashtags turned out to be problematic. Collection of pro-vaccination tweets did not present any difficulties (with hashtags like *#vaccineswork*, *#vaccinateUS*, *#vaccinateYourKids*), however collecting data with specific anti-vaccination content turned out to be more challenging. The tweets collected with supposedly anti-vaccination hashtags (*#antivax*, *#antivaxxers*, *#dropvaccine*, *#Vaccine-hesitancy*) contained both favour and against stances with

a heavy shift towards the former (i.e. these hashtags appear to be used in pro-vaccination rhetoric when referring to anti-vaccination movement, but not by the anti-vaccination advocates themselves).

Eventually, we decided that the most appropriate approach for our case was to search tweets based on stance-neutral hashtags like *#vaccine* and *#vaccination*. This approach was prompted by a study on two twitter datasets by Blankenship et al [1].

A rough preliminary annotation effort for collected tweets has revealed a more balanced distribution of stance:

- *#vaccine* dataset: out of 100 pre-annotated tweets 26 are against, 18 pro, 33 neutral and 8 misc*[1] tweets;
- *#vaccination* dataset: out of 30 pre-annotated tweets 6 are against, 12 are neutral, 10 are pro, 4 are misc.

A research into the vaccination debate and hashtags more often used by tweeters who used other popular vaccination-related hashtags has also revealed the usage of hashtags *#unvaxxed* and *#vaxxed*; a preliminary annotation has shown a promising distribution of stance towards vaccination debate.

- *#unvaxxed* dataset: out of 36 pre-annotated tweets 24 are against, 4 are neutral, 6 are pro, 2 are misc;
- *#vaxxed* dataset: out of 30 pre-annotated tweets 13 are against, are neutral, 8 are pro, 4 are misc.

3.3 Data annotation framework

Preliminary analysis of tweets collected has revealed recurring topics and rhetoric: e.g., measles (and other preventable deceases) outbreaks, pro-vaccination and anti-vaccination legislation, personal accounts (pro and against) of peoples' experience with vaccination, new advances in vaccination research. Some of the rhetoric directions present across the dataset are as follows: criticism of governmental bodies, criticism of antivaccination movement, exposure and refutation of allegedly untrue statements (both pro and against). This study of linguistic/topical peculiarities has allowed us to form the approach to defining stance labels, introduce additional features and venture into exploring the data we have collected.

Stance feature is the main point of interest and the central classification task of the project. The overview of the stance labels and their definitions are as follows:

FAVOR: tweeter expresses trust/positive attitude towards usage and/or effectiveness of vaccines, vaccination and aspects of it; tweeter expresses support of negative reaction and criticism towards anti-vaccination movement; the tweet depicts vaccination efforts in positive light.

- *According to the @WHO, #measles cases jumped in Europe, where 35 people have died from the #disease. Although highly contagious, measles is preventable with a #vaccine*
- *Nevertheless the #vaccine will – and should – be used in the next #Ebola outbreak, especially on 1st responders & hlth care wrkers.*
- *Note the well-explained description of #herdimmunity: It doesn't mean disease will never appear or spread. Rather, it means w fewer hosts, disease is less likely to become outbreak/#epidemic. So yes, your #unvaxxed family is a danger to others. #antivax*

[1]Miscellaneous category was dropped for the final annotations

AGAINST: tweeter expresses distrust of vaccines and/or their effectiveness; states danger of/expresses fear of adverse effects; expresses distrust and criticism of authorities/government action in relation to vaccines/vaccination.

- *Although pediatricians have a legal duty to fully inform patients about #vaccine risks and side effects, the lure of monetary perks and the desire to fit in may lessen their motivation to do so*
- *Ha ha - business as usual. See 6-step #vaccine scam recipe at the bottom*
- *#Hepatitis B is a sexually transmitted disease, and infants from mothers who don't have HepB (over 98% in the US) are not at risk of developing it at all. It means that for 98% of babies in the US HepB #vaccination poses no benefits, only risks.*

NEUTRAL: tweeter states facts about vaccination figures/statistics without giving opinion; tweeter describes situations related to vaccination context without personal perspective.

- *Italy bans unvaccinated children from school. https:...*
- *New work in the March issue examines why #influenza #vaccination elicits poor efficacy in elderly individuals & uncovers a reduced accumulation of de novo immunoglobulin gene somatic mutations & poorly adapted #antibody responses upon #flu vaccination*
- *Hundreds converge on #Maine capital for #vaccination bill hearing http://... #mepolitics*

In addition to stance feature, which is the main focus of our classification task, we have added the following features, which reflect the characteristics of our dataset:

- Fact: Initial study of the dataset has revealed that a number of tweets do not contain a personal element/opinion, but only facts about vaccination (without any opinion or personal reflection involved). We decided to label these tweets as a separate attribute, to see if this plays any role or has any consequence for active learner framework. An example of a fact tweet is as follows:
  Demand for #measles #vaccine has surged in the Washington county where the virus is linked to more than 50 cases: up nearly 500% percent in January compared to the previous January
- Stance annotation confidence score (0 to 5): The nature of the tweets and communication on Twitter platform factor into varying annotation confidence: the debate around vaccination demands some knowledge of context. Moreover, irrespective of topic, tweets often appear unclear and difficult for interpretation. In order to measure this quality of uncertainty we have introduced a 0 to 5 confidence score for labeling stance. No annotation rules were established for confidence, since we wanted this to be a subjective indicator of the certainty of an annotation, in contrast to the more objective indicator of inter-annotate agreement.

3.4 Active Learning experimental setup

To test the effectiveness of the active learning strategies, we designed an experiment in which classifiers are trained multiple times with different active learning strategies. By incrementally increasing the size of the training corpus, we can monitor the learning

	Low confidence (0-2)
Favor	Apparently Kelly's polio is acting up again! Not vaccinated, not allowed in public so says the Bible! *if they can make shit up, so can I #Vaccine #TreasonousGOP #FakeChristians #LGBTQ https://...
Against	Pharma's tarnished reputation helps fuel the anti-#vaccine movement
Neutral	#Vaccine Panav Bio-Tech Introduces Classical Swine Fever

	Medium confidence (3)
Favor	Conference: Malaria Vaccines for the World - MVW 2019, 8-10 May 2019 University of Oxford, Oxford, UK #malaria #vaccine #conference
Against	#MandatoryVaccines are #NationalSecurity risk since evil #NWO #NaZi #Monsters can and do cause #Autism,#LearningDisAbility,#ImmunityDisOrder and even #Murder entire generations of #Goyim by purposefully corrupting just 1 mandatory #Vaccine
Neutral	Structural studies of a #vaccine protein antigen reveal ability to generate cross-reactive #antibodies. #crystallography #immunology https://...

	High confidence (4-5)
Favor	The #vaccine that protects against HPV, #Gardasil 9 is now approved for men and women up to age 45. Look into it: it's a safe and effective way to protect against several types of cancer that you do not want to get.
Against	#Vaccine #Safety Commission: 50 #Studies. Vaccine studies vaccine safety agencies neglected to mention. https://...
Neutral	Findings from a systematic review of published studies showed there is no need for an additional dose of #measles #vaccine in HIV-infected adolescents and adults. Read more: https://...

rate, and compare this for the different strategies. Specifically, the following pipeline was used:

(1) *Preprocessing:* we preprocessed the data by removing unuseful information like stop words and by replacing unique identifiers (e.g., names, urls) with fixed labels. Users' mentions were replaced with the word 'mention' and link to other websites were replaced with the words *'url'* or *'twitter picture'* if the link was a link to a twitter's image. Finally, the number sign # was removed from the hashtags, and some of them were expanded when composed by more than one word, like *'vaccinesworks'* into *'vaccines works'*.

(2) *Corpus split:* We then split the corpus into three datasets, one for training and validation, one for testing and one for the active learning sampling. Specifically, we used the set of 400 jointly annotated tweets for the active learning sampling, since we were interested in checking if any relationship between the confidence and the active selection could be found. The remaining tweets were split into 500 instances for training and validation, and 100 (+20) instances for testing.

(3) *Models:* We implemented using PyTorch [12] the multichannel CNN proposed in [6] and manually tuned the parameters in order to avoid overfitting as much as possible. We find out that for our dataset a less complex architecture with two channels of size 2 and 3 (bigrams and trigrams) and 50 filters works better than the original configuration used in Kim's paper. The model was trained using the pre-trained GloVe embedding vectors [14] of size 50. As a baseline, we also tested a Naive Bayes classifier using the MultinomialNB implementation available in sklearn [13]. The Naive Bayes classifier was trained by applying to the preprocessed data a TF-IDF vectorizer.

(4) *Active learning strategies:* For the Multichannel CNN we tested 4 strategies, namely the random selection, the entropy uncertainty sampling, the cost-effective active learning framework and the deep active learning by dropout disagreement. For the Naive Bayes classifier we tested also 4 strategies, namely the above-mentioned random selection and entropy uncertainty sampling, and other two uncertainty strategies, the max-margin and least confident ones.

(5) *Simulations:* We performed 30 simulations for each active learning strategy by training an initial model on the training and validation dataset (randomly split into training and validation for each simulation) and by iteratively retraining the model adding tweets selected with each strategy. All simulations were performed two times, one using a sample size of 25 tweets for each active selection and one using a sample size of 50, for a total of 240 simulations.

4 RESULTS

We will first present the inter annotator agreement scores obtained for the annotations of our corpus. Then we will present and discuss the results of the active learning experiment described in the previous section.

4.1 Inter-annotator agreement

The Inter-annotator agreement (IAA) shows how strongly the four annotators agreed on the stance labels. Given the prior training on this annotation task, this measure provides a good indication of how difficult it is, even for human coders, to obtain a reliable classification measure. Beside the IAA of the stance labels, we also report agreement on annotating the *fact* feature and the confidence agreement. Although these labeling tasks were not trained on, they provide additional insight into the complications of annotating stances.

(1) The overall agreement between all four annotators
(2) Pairwise Agreement between each pair of annotators
(3) Agreement on *fact* feature
(4) Confidence agreement

The figures (as presented in table 5) demonstrate that the total agreement on stance reaches a 0.61, signifying moderate level of consistency in terms of determining stance, with *fact* category gaining slightly lower agreement. Overall, this shows that while there is clearly a common interpretation of stance among the annotators, the agreement is far from perfect.

A close look revealed that maximum agreement was achieved in cases when the tweet clearly and directly stated that vaccines

are a benefit (or otherwise, a danger), or, if a tweet carried mostly informational value, yet still expressed positive attitude towards vaccine through declaring it's usefulness; such as in cases below:

- Getting a #flu vaccine has never been more convenient - thanks to pharmacies like @riteaid - and it's also the single best way to protect yourself against the dangers of #flu. What are you waiting for? Get a #flu #vaccine today in honor of #NIVW
- "Dr. Angie Myers, @ChildrensMercy shared a wealth of information today about the human papillomavirus (#HPV), the low #vaccine rates in the metro area, and what is being done to protect more citizens from this dangerous, potentially #cancer-causing disease.

A case below, on the other hand, presents a disagreement between the annotators because no direct evaluation is made, when compared to cases above, and making a judgment requires a deeper knowledge of context and debate tropes; in this particular case - of the fact that anti-vaccination movement supporters often employ the argument of parental rights (in the sense of having the right to refuse to vaccinate one's child).

- Exciting conversations with a #California school of #socialwork, a #vaccine apologist, and a government! We're #GettingThrough for #ParentalRights

Moreover, if we look into binary agreement on annotating in terms of "belonging to the category/not belonging to the category", as illustrated by table 7, we can see that annotating *neutral* stance has presented the biggest difficulty for the annotators since the agreement for annotating neutral category as opposed to *favour* and *against* drops considerably. This might partly be attributed to the fact that a fair share of tweets reveal the stance in a non-explicit way, and require knowledge of context to make judgement about a tweet. Neutral tweets also often find themselves on a borderline between *favour* and *neutral* category since there is a considerable amount of tweets that use non-expressive and non-evaluative language to express a *favour* stance in an implicit way.

The IAA agreement on *fact* feature demonstrates a lower score (0.48), and after a detailed analysis this might be attested to the subtlety of opinion expressed in tweets that caused a disagreement. The tweet below creates a disagreement most likely because of the different treatment of the quote: while some of the annotators considered it purely part of the announcement, others treated it as showing implicit approval:

- @RichClarkePsy is discussing the #Vaccine Confidence Project: Across the European Union, vaccine delays and refusals are contributing to declining immunisation rates in a number of countries and are leading to increases in disease outbreaks.

Another point of disagreement is demonstrated in the following tweet, where different impression of the statement might have contributed to different labels assigned; the opinion element is very slight, but could be treated as hinting at disapproving attitude:

- Anti-#vaccine Italian government fires entire health expert board

Compared to the tweet below, which was considered to contain a fact by all 4 annotators, since it does not contain any hints to being even slightly opinionated:

Table 5: Overall IAA

Annotation	Score	Statistic
Stance	0.61	Fleiss K
Contains a fact	0.48	Fleiss K
Confidence	0.26	Krippendorff α

Table 6: Stance pairwise IAA (Cohen K)

Annotator	A1	A2	A3	A4
A1		.59	.58	.63
A2	.59		.56	.69
A3	.58	.56		.63
A4	.63	.69	.63	

- Structural studies of a #vaccine protein antigen reveal ability to generate cross-reactive #antibodies

Table 7: Binary stance annotations (Fleiss k score)

Category	Score
favor / not favour	0.7
against / not against	0.7
neutral / not neutral	0.3

4.2 Active learning experiment: Convolutional Neural Network

We reported in Fig. 1 the learning rates, intended here as the rate of performance increasing after each retraining, of the different AL strategies used with the CNN. The black dotted lines represent the performance of the CNN trained with the whole data without active learning selection. Using the whole dataset the maximum accuracy achieved is 52.4% while the maximum f-score is 43.1%. This is close to the reported accuracy (60.5%) of a recent work in which the authors also trained a CNN with GloVe word embedding in a stance classification task, but on italian tweets [2].

All the strategies reach better scores, including the random selection one. This might be due to the fact that for the model trained on the whole data we used a callback function as a criteria to stop the training, while for the models trained with active learning we set a fixed amount of training epochs (5) for each retraining to decrease the total amount of time required to run each simulation. This means that these models were trained for 45 epochs in total, while the whole model took on average 20 epochs to stop training because of the callback function. The difference in the number of training epochs might explain the small gain observed also for the random selection.

Among the tried strategies, CEAL is the one that reaches the higher scores for both, accuracy: 55.0% and f-score: 47.0%. The second best results are the Entropy ones, slightly outperforming the random selection at every active step like CEAL. Since the CEAL strategy ranks the unlabelled data based on their predictions' entropy as the Entropy selection strategy, and since the latter shows a slightly lower learning rate, we can say that our results support

the underlying idea of the CEAL strategy of leveraging automatic annotations generate by the model.

Another important aspect to point out is that the DO-BALD strategy was not able to outperform the random selection, reaching at the end the same accuracy but lower f-score. This might be due to the fact that the CNN we implemented includes only 3 dropout layers, causing an overall small amount of variation in the predictions for each unlabelled instance when performing the monte-carlo sampling for the active selection. This hypothesis is supported by a deeper inspection of the average variance estimated for each tweet in the set of unlabelled data. At each active step, the difference between the maximum variance and the minimum variance was indeed always in the order of 10^{-2}.

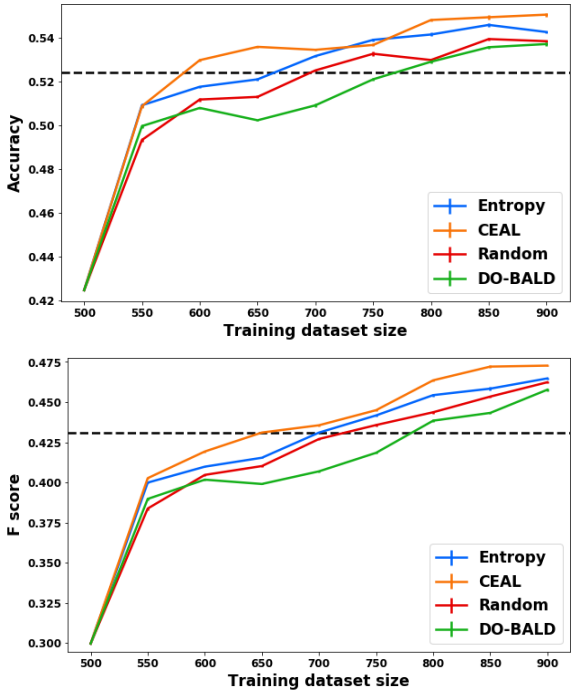

Figure 1: Comparison of accuracy and F1 score after each retraining for each active learning strategy. The black dotted lines represents the scores of the classifier trained with all available data (900 tweets) at once.

4.3 Active learning experiment: Naive Bayes

Fig2 shows the results for the active learning experiment performed with the NB classifier trained with the TF-IDF representation of the tweets tokens. In this case, all the strategies reach the same accuracy and f-score at the end. This is expected since unlike deep learning models, the NB does not use any weights update rule or loss function, but it just estimates the joint probability distribution between the input features and the target labels. However, the max margin sampling strategy proved to be most effective in the selection of relevant tweets from the set of unlabelled data, since it managed to reach the best possible accuracy and f-score in 5 active learning steps out of 8, i.e. requiring 150 tweets less than other strategies.

Unlikely for the CNN, the Entropy selection did not outperform the random selection, this is also true for the least confident strategy.

As for the CNN, the incremental learning shows to reach higher performances than the model trained on the whole dataset at once. This time, the blame probably goes to the tf-idf encoding of the tweets words tokens. Because of how the scikit-learn implementation of the TF-IDF vectorizer works, one cannot update the vectorizer dictionary at each retraining, i.e. the Naive Bayes is retrained by updating only the frequency of the words that were presented during the initial training of the model. Words which appear only in tweets selected during the active learning steps are not used as new features. Instead, when training the model with the whole dataset, the vectorizer encodes all the words which appear in the tweets, generating a bigger number of features. It seems that this forced features pruning bring better generalization of the NB than using the whole vocabulary of our corpus.

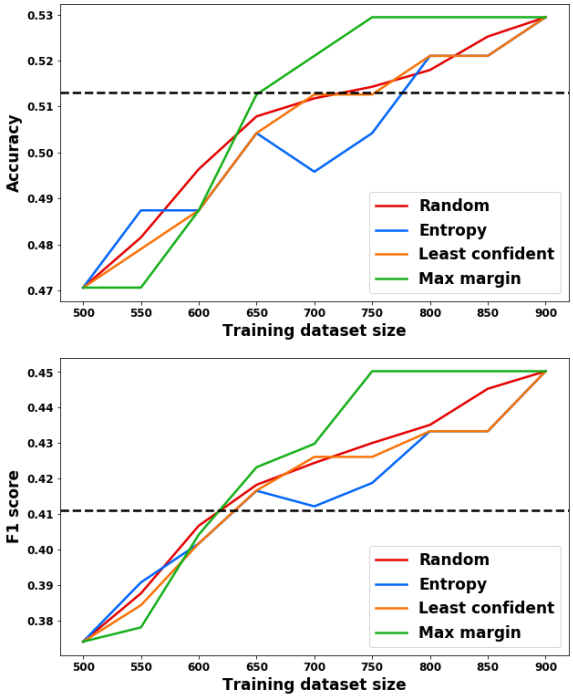

Figure 2: Comparison of accuracy and F1 score after each retraining for Multinomial Naive Bayes baseline. The black dotted lines represents the scores of the classifier trained with all available data (900 tweets) at once.

4.4 Confidence and Active selection comparison

In this paper, we were interested in checking if in our experiment, the active learning strategies would have selected at each retraining step tweets which were not only ranked as interesting by the model,

i.e. hard to classify, but tweets which were also hard to annotate for us during the creation of the corpus.

As already explained in the previous sections, to test this hypothesis we annotate 400 tweets with a confidence score, and then we averaged our single confidence scores in order to obtain a unique final mean confidence score for each tweet. These tweets were used as the dataset of unlabelled data for the active learning sampling during our experiment. Since we performed multiple simulations for each strategy, we saved during each simulation every set of tweets selected at every active learning step. In this way, we were able to count in how many simulations a specific tweet was selected during a specific active learning step. Of course, during each simulation each tweet could have been selected only once at a specific step, so we expected tweets with a low mean confidence score to be selected mostly during the initial active steps. The graphs with a unique line in Fig3 show for each strategy (applied with the CNN) the average confidence of the tweets selected at each active steps. Here, average score refers to a weighted average in which each tweet's mean confidence score was multiplied by the number of times that specific tweet was selected in a specific active step over the 30 simulations. The resulting values were then summed and divided by 1500 (50 -selected tweets per step- * 30 -number of simulations-).

We can see that the two strategies that performed better than the random selection, i.e. CEAL and Entropy, show an increasing trend in terms of confidence of the selected tweets, meaning that these strategies indeed selected tweets with a lower confidence in the first steps. The random selection instead selected tweets with a constant average confidence as well as the DO-BALD strategy.

In Fig3 we also reported the average amount, in percentage, of selected tweets per stance at each active step. For the random selection, we can see that the percentage of selected tweets per stance remain constant over each active step, as expected from a random sampling. The 3 flat lines in the bottom right graph indeed represent exactly the same percentages of against, favor and neutral tweets in the dataset used to perform the active selection. Specifically, the dataset contains 49.75% of against tweets, 31.75% of favor tweets and 18.50% of neutral tweets, exactly the values of the flat lines.

Again, we can see a similar trend for the Entropy and CEAL strategies. The interesting thing to stress out here is the constant decrease in the selection of neutral tweets. Both strategies collect in the first steps more neutral and against tweets and less favor tweets than the random selection. The trend for the selection of favor and against tweets then change over the active steps, while for the neutral tweets the trend is linear, suggesting that these strategies tried to collect a bigger amount of neutral tweets in the first active steps. This is an interesting fact because the average mean confidence score of the neutral tweets in the dataset was 2.89, significantly lower than the average mean confidence score of favor tweets: 3.90 and against tweets: 3.64.

Different selection trends can be seen for the DO-BALD strategy. Fewer favor tweets were selected compared to the Entropy and CEAL strategies, but against tweets were preferred over the neutral once in the first steps. This might be another clue to explain why

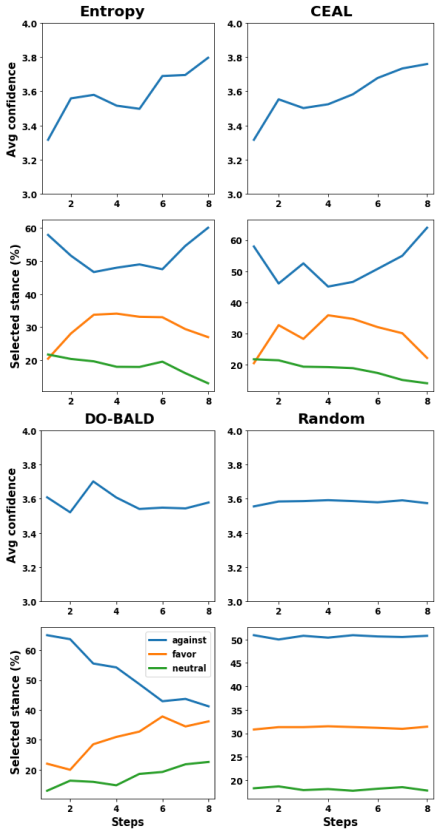

Figure 3: Comparison of average confidence and stance percentage of the selected tweets at each active learning steps.

the DO-BALD strategy performed on overall worse than all the others.

5 CONCLUSION

In this paper we investigated whether using an active learning framework can be effective in reducing the required annotated data for training a good stance detection model. Our results showed that active learning provides only a small benefit on the overall performances of the classifiers trained for our specific stance detection task, without significantly decreasing the number of data required to reach the highest accuracy and f-score achievable with our corpus.

The results showed also that there was no significant difference between traditional active learning pool-based uncertainty sampling strategies applied to the shallow model NB and more recent active learning strategies proposed for deep models applied to a CNN. We must point out anyway that the NB was trained using only basic features. Including syntactic and semantic features like n-grams or dependencies will most likely boost the performance of the classifier, make it preferable to the CNN. Nevertheless, in our paper we used the NB only as a baseline for the CNN model, so we leave as a future work the possibility to deeply analyze the impact

of active learning in combination with the use of different sets of linguistic features.

Another hypothesis we wanted to test was the possibility to find relationships between the tweets selected with the active learning strategies and our annotations, specifically confidence and stance. We tested if this was the case using the tweets selected by the CNN and we found that two strategies, Entropy and CEAL, give priority to tweets with low average confidence when performing the active selection. The same strategies also showed interesting patterns when we analyzed the percentage of tweets selected per stance. Neutral and favor tweets are preferred over the against ones in the first active steps. These two patterns, i.e. the selection of low confidence tweets and the selection of more neutral and favor tweets during the first active steps are strictly connected, since in our corpus favor and neutral tweets were annotated with lower confidence by all the annotators. This is an interesting result that should be further investigated, for example by expanding the dataset for the active learning sampling in order to make it balanced. With such a dataset, the amount of neutral tweets selected by the active strategies in the first steps should increase. If this is the case, the active learning framework could be turned into a tool not meant to reduce the amount of data to annotate, but meant to help annotators to create a balanced corpora in a more efficient way than simply annotating as much documents as possible and then discarding part of them.

The code implementation for this paper is available on github [2]. Data are also available upon request to one of the authors.

REFERENCES

[1] Blankenship, E. B., Goff, M. E., Yin, J., Tse, Z. T. H., Fu, K.-W., Liang, H., Saroha, N., and Fung, I. C.-H. (2018). Sentiment, contents, and retweets: A study of two vaccine-related twitter datasets. *The Permanente journal*, 22.

[2] D'Andrea, E., Ducange, P., Bechini, A., Renda, A., and Marcelloni, F. (2019). Monitoring the public opinion about the vaccination topic from tweets analysis. *Expert Systems with Applications*, 116:209 – 226.

[3] Fedorov, V. V. (2013). *Theory of optimal experiments*. Elsevier.

[4] Gal, Y. and Ghahramani, Z. (2016). Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059.

[5] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1321–1330. JMLR. org.

[6] Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*.

[7] Lewis, D. D. and Gale, W. A. (1994). A sequential algorithm for training text classifiers. In *SIGIR'94*, pages 3–12. Springer.

[8] Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

[9] Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., and Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150, Portland, Oregon, USA. Association for Computational Linguistics.

[10] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X., and Cherry, C. (2016a). Semeval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41.

[11] Mohammad, S., Kiritchenko, S., Sobhani, P., Zhu, X.-D., and Cherry, C. (2016b). A dataset for detecting stance in tweets. In *LREC*.

[12] Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., and Lerer, A. (2017). Automatic differentiation in pytorch. In *NIPS-W*.

[13] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

[14] Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

[15] Settles, B. (2012). Active learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 6(1):1–114.

[16] Seung, H. S., Opper, M., and Sompolinsky, H. (1992). Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM.

[17] Shannon, C. E. (1948). A mathematical theory of communication. *Bell system technical journal*, 27(3):379–423.

[18] Siddhant, A. and Lipton, Z. C. (2018). Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. *arXiv preprint arXiv:1808.05697*.

[19] Somasundaran, S. and Wiebe, J. (2010). Recognizing stances in ideological online debates. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 116–124. Association for Computational Linguistics.

[20] Thomas, M., Pang, B., and Lee, L. (2006). Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics.

[21] Walker, M. A., Anand, P., Abbott, R., Tree, J. E. F., Martell, C., and King, J. (2012). That is your evidence?: Classifying stance in online political debate. *Decision Support Systems*, 53(4):719–729.

[22] Wang, K., Zhang, D., Li, Y., Zhang, R., and Lin, L. (2016). Cost-effective active learning for deep image classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(12):2591–2600.

[23] Xiao, Y. and Wang, W. Y. (2018). Quantifying uncertainties in natural language processing tasks. *arXiv preprint arXiv:1811.07253*.

[24] Young, T., Hazarika, D., Poria, S., and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine*, 13(3):55–75.

[25] Zhang, L., Wang, S., and Liu, B. (2018). Deep learning for sentiment analysis: A survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1253.

[2] https://github.com/EdoardoGuerriero/Building-task-specific-stance-identification-models-anevaluation-of-the-active-learning-approach