Joint Rumour Stance and Veracity

Anders Edelbo Lillie[*]
ITU Copenhagen
Denmark
aedl@itu.dk

Emil Refsgaard Middelboe[*]
ITU Copenhagen
Denmark
erem@itu.dk

Leon Derczynski
ITU Copenhagen
Denmark
ld@itu.dk

Abstract

The net is rife with rumours that spread through microblogs and social media. Not all the claims in these can be verified. However, recent work has shown that the stances alone that commenters take toward claims can be sufficiently good indicators of claim veracity, using e.g. an HMM that takes conversational stance sequences as the only input. Existing results are monolingual (English) and mono-platform (Twitter). This paper introduces a stance-annotated Reddit dataset for the Danish language, and describes various implementations of stance classification models. Of these, a Linear SVM provides predicts stance best, with 0.76 accuracy / 0.42 macro F_1. Stance labels are then used to predict veracity across platforms and also across languages, training on conversations held in one language and using the model on conversations held in another. In our experiments, monolinugal scores reach stance-based veracity accuracy of 0.83 (F_1 0.68); applying the model across languages predicts veracity of claims with an accuracy of 0.82 (F_1 0.67). This demonstrates the surprising and powerful viability of transferring stance-based veracity prediction across languages.

1 Introduction

Social media has come to play a big role in our everyday lives as we use it to connect with our social network, but also to connect with the world. It is common to catch up on news through Facebook, or to be alerted with emerging events through Twitter. However these phenomena create a platform for the spread of rumours, that is, stories with unverified claims, which may or may not be true (Huang et al., 2015). This has lead to the concept of *fake news*, or misinformation, where the spreading of a misleading rumour is intentional (Shu et al., 2017). Can we somehow automatically predict the veracity of rumours? Research has tried to tackle this problem (Qazvinian et al., 2011), but automated rumour veracity prediction is still maturing (Gorrell et al., 2019).

This project investigates stance classification as a step for automatically determining the veracity of a rumour. Previous research has shown that the stance of a crowd is a strong indicator for veracity (Dungs et al., 2018), but that it is a difficult task to build a reliable classifier (Derczynski et al., 2017). Moreover a study has shown that careful feature engineering can have substantial influence on the accuracy of a classifier (Aker et al., 2017). A system able to verify or refute rumours is typically made up of four components: rumour detection, rumour tracking, stance classification, and veracity classification (Zubiaga et al., 2018). This project will mainly be concerned with stance classification and rumour veracity classification.

Current research is mostly concerned with the English language, and in particular data from Twitter is used as data source because of its availability and relevant news content (Derczynski et al., 2017; Gorrell et al., 2019). To our knowledge no research within this area has been carried out in a Danish context. To perform automated rumour veracity prediction for the Danish language following the components in Zubiaga et al. (2018), a number of problems must be solved. (1) to facilitate Danish stance classification a Danish dataset must be generated and annotated for stance. (2) developing a good stance classifier is difficult, especially given the unknown domain of the Danish language. Therefore experiments must be per-

∗: These authors contributed to the paper equally.

Proceedings of the 22nd Nordic Conference on Computational Linguistics (NoDaLiDa), pages 208–221

formed to investigate what approach to apply to Danish stance classification. (3) given rumourous data, and aided by stance classification, a rumour veracity prediction component should be able to determine whether it is true or false.

2 Background

The attitude that people express towards claims can be used to predict veracity of those claims, and these attitudes can be modelled by stance classifiers. This section will cover some state-of-the-art research for stance classification and rumour veracity resolution. While the introduced classification tasks are related to the work carried out in this project, they differ in a number of ways: (1) this project performs stance classification for the Danish language, (2) the generated dataset is from the Reddit platform, and (3) this project seeks to join stance classification and veracity prediction without external fact verification.

Stance classification: Long-Short Term Memory (LSTM) neural network models are popular, as they have proven to be efficient for working with data within NLP. In particular (Kochkina et al., 2017) introduced a stance classifier based on a "Branch-LSTM" architecture: instead of considering a single tweet in isolation, whole branches are used as input to the classifier, capturing structural information of the conversation. The model is configured with several dense ReLU layers, a 50% dropout layer, and a softmax output layer, scoring a 0.78 in accuracy and 0.43 macro F_1 score. They are however unable to predict the under-represented "denying" class.

Another LSTM approach deals with the problem introduced in the SemEval 2016 task 6 (Mohammad et al., 2016). The LSTM implements a bidirectional conditional structure, which classifies stance towards a target with the labels "positive", "negative", and "neutral" (Augenstein et al., 2016). The approach is unsupervised, i.e. data is not labelled for the test targets in the training set. In this case the system achieves state-of-the-art performance with a macro F_1 score of 0.49, and further 0.58 when applying weak supervision.

A different approach is based on having well-engineered features for stance classification experiments using non-neural networks classifiers instead of Deep Learning (DL) methods (Aker et al., 2017). Common features such as CBOW and POS tagging are implemented, but are extended with problem-specific features, which are designed to capture how users react to tweets and express confidence in them. A Random Forest classifier performed best, with an accuracy of 0.79.

The lack of labelled data is a major challenge for stance classification. One study shows that classification can be improved by transferring knowledge from other datasets (Xu et al., 2019). In particular, a model is implemented with *adversarial domain adaptation* to train on the FEVER dataset (Thorne et al., 2018) and test on the Fake News Challenge dataset.[1] By augmenting the traditional approach for stance classification with a domain adaption component, the model learns to predict which domain features originate from.

RumourEval 2019 is a very recent SemEval task which deals with stance classification and veracity prediction (Gorrell et al., 2019), and a first look at the scoreboard indicates very promising results.[2] With the Branch-LSTM approach as a baseline on the RumourEval 2019 dataset, scoring 0.4930 macro F_1, the "BERT" system scores a macro F_1 of 0.6167 (Fajcik et al., 2019). The implementation employs transfer learning on large English corpora, then an encoding scheme concatenates the embeddings of the source, previous and target post. Finally the output is fed through two dense layers to provide class probabilities. These BERT models are used in several different ensemble methods where the average class distribution is used as the final prediction.

Rumour veracity prediction: Rumour veracity classification is considered a challenging task as one must typically predict a truth value from a single text, being the one that initiates the rumour. The best performing team for that task in RumourEval 2017 (Derczynski et al., 2017) implements a Linear Support Vector Machine (SVM) with only few (useful) features (Enayet and El-Beltagy, 2017). They experiment with several common features such as hashtag existence, URL existence, and sentiment, but also incorporates an interesting feature of capturing whether a text is a question or not. Furthermore the percentage of replying tweets classified as supporting, denying, or querying from stance classification is applied. It is concluded that content and Twitter features were the most useful for the veracity classification task

[1] http://www.fakenewschallenge.org

[2] https://competitions.codalab.org/competitions/19938 26-05-2019

and the system scores an accuracy of 0.53.

While the system described above engages in the task of resolving veracity given a single rumour text, another interesting approach is based on the use of crowd/collective stance, which is the set of stances over a conversation (Dungs et al., 2018). This system predicts the veracity of a rumour, based solely on crowd stance as well as tweet times. A Hidden Markov Model (HMM) is implemented, which is utilised such that individual stances over a rumour's lifetime is regarded as an ordered sequence of observations. This is then used to compare sequence occurrence probabilities for true and false rumours respectively. The best scoring model, which include both stance labels and tweet times, scores an F_1 of 0.804, while the HMM with only stance labels scores 0.756 F_1. The use of automatic stance labels from (Aker et al., 2017) is also applied, which does not change performance much, proving the method to have practical applications. It is also shown that using the model for rumour veracity prediction is still useful when limiting the number of tweets to e.g. 5 and 10 tweets respectively.

Danish: While Danish is not a privileged language in terms of resources (Kirkedal et al., 2019), there is stance classification work on political quotes (Lehmann and Derczynski, 2019). However, this is over a different text genre, and does not focus on veracity prediction as its final goal.

Comprehensive reviews of automatic veracity and rumour analysis from an NLP perspective include Zubiaga et al. (2018), Atanasova et al. (2019), and Lillie and Middelboe (2019b).

3 Dataset

Because of various limitations on big social media platforms including Facebook and Twitter, Reddit is used as platform for the dataset.[3] This is a novel approach; prior research has typically relied on Twitter (Mohammad et al., 2016; Derczynski et al., 2017; Gorrell et al., 2019).

Data sampling: The data gathering process consists of two approaches: to manually identify interesting submissions on Reddit, and; to issue queries to the Reddit API[4] on specific topics. An example of a topic could be "Peter Madsen" referring to the submarine murder case, starting from August 2017.[5] A query would as such be constructed of the topic "Peter Madsen" as search text, a time window and a minimum amount of Reddit upvotes. A minimum-upvotes filter is applied to limit the amount of data returned by the query. Moreover the temporal filters are to ensure a certain amount of relevance to the case, specifically *when* the event initially unfolded. Several submissions prior or subsequent to the given case may match a search term such as "ubåd" (submarine).

Four Danish Subreddits were browsed, including "Denmark, denmark2, DKpol, and GammelDansk",[6] although all relevant data turned out to be from the "Denmark" Subreddit. The submission IDs found manually and returned by the queries are used to download all posts from each submission using the `praw`[7] and `psaw`[8] Python libraries. The submission data is subsequently stored in a JSON format, one JSON file per submission, consisting of submission data and a list of comment data. These files include submission post text and comment text, as well as meta-information about the following: submission post, submitter user, Subreddit, comments, and commenting users.

Annotation: One widely used annotation scheme for stance on Twitter is the SDQC approach from Zubiaga et al. (2016). Twitter differs from Reddit in the way conversations are structured. Each tweet spawns a *conversation* which can have nested replies, and as such creates *branches*. Reddit implements the same mechanism, but a set of conversations are tied to a specific submission, which is initiated by a submission post. The Reddit structure is depicted in Figure 1, illustrating a conversation (in green) and two respective branches (in respectively red and purple). Note that branches share at least one comment. Thus, a way to annotate data from the Reddit platform with the annotation scheme from Zubiaga et al. (2016) is by regarding a submission post as a source, instead of each top-level comment for the conversations.

The stance of the source/submission post is taken into account when annotating the stance for

[3] In particular the Danish Subreddit at `www.reddit.com/r/Denmark/`

[4] `www.reddit.com/dev/api/`

[5] `www.dr.dk/nyheder/tema/ubaadssagen`

[6] `www.reddit.com/r/Denmark/wiki/danish-subreddits`

[7] `praw.readthedocs.io/en/latest/` v. 6.0.0.

[8] `github.com/dmarx/psaw` v. 0.0.7.

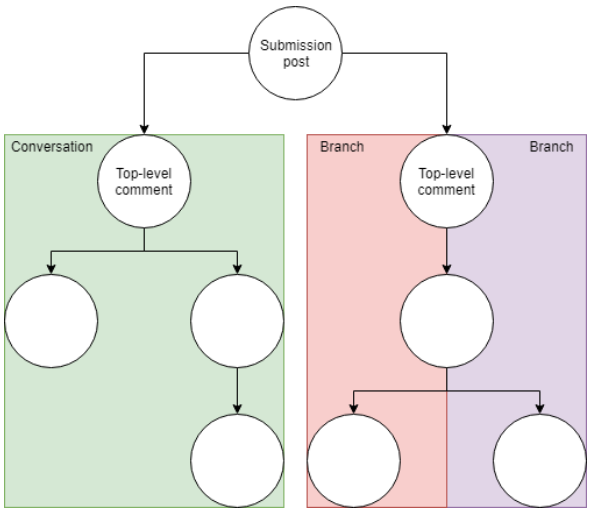

Figure 1: The structure of a Reddit submission

replying posts of top-level posts. As stance annotations are relative to some target, each post does not have one single stance annotation: each post is annotated for the stance targeted towards the submission and the stance targeted towards the direct parent of the post. The double-annotation should facilitate a way to infer the stance for individual posts. For instance, if the source post supports a rumour, and a nested reply supports its parent post, which in turn denies the source, then the nested reply is implicitly denying the rumour.

Further, a majority of submissions have no text, but a title and a link to an article, image or another website, with content related to the title of the submission. If this is the case and the title of the submission bears no significant stance, it is assumed that the author of the submission takes the same stance as the content which is attached to the submission.

Annotation tool: A custom web-based annotation tool was built to facilitate the annotation process of the data. C# and MySQL technologies were used to build the tool in order to support rapid development. The tool enables annotators to partition datasets into events and upload submissions in the JSON form from the gathering Reddit data to each event. Further the tool allows for a branch view of each submission in the event and facilitates annotation following the SDQC scheme, as well as certainty and evidentiality as presented by Zubiaga et al. (2016). Any annotation conflicts are highlighted by the tool, which will cause the annotators to discuss and re-annotate the post with a conflict. A screenshot of the annotation page for

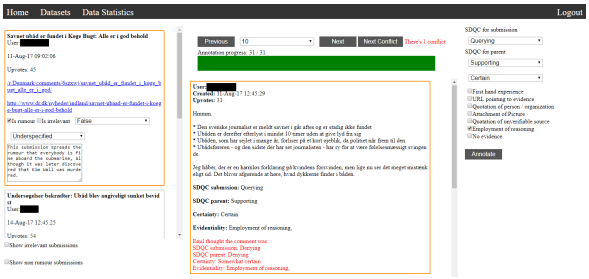

Figure 2: Screenshot of the annotation tool

the annotation tool is presented in Figure 2.

During annotation of the first $\sim$ 500 posts, annotators disagreed upon labels for around 40-50% of posts. However after the initial annotation work this rate dropped to around 25%. Annotation conflicts were handled in collaboration between the annotators after annotation of every $\sim$ 100 posts.

DAST: The result of the data sampling and annotation process is the **Da**nish **st**ance-annotated Reddit dataset (DAST). The dataset consists of a total of 11 events with 3,007 stance-annotated Reddit posts across 33 submissions with a total of 1,161 branches. Information on DAST is presented in Table 1 including event names, SDQC distribution and total post counts.

Label / *Event*	S	D	Q	C	*Total*
5G	26	47	7	193	273
Donald Trump	39	17	5	185	246
HPV vaccine	24	4	8	219	255
ISIS	3	40	8	118	169
"Kost"	50	56	4	447	557
MeToo	1	8	3	48	60
"Overvågning"	41	20	13	278	352
Peter Madsen	15	45	19	302	381
"Politik"	43	46	7	227	323
"Togstrejke"	8	6	3	84	101
"Ulve i DK"	23	11	4	252	290
Total	273	300	81	2,353	3,007
%	9.1	10	2.7	78.2	100

Table 1: SDQC stance labels per event

The "querying" label is rare with a total of 81 annotations out of the 3,007 posts. The "supporting" and "denying" labels are almost equally distributed with a total of respectively 273 "supporting" and "300" denying posts. The "commenting" class is the absolute dominant one, with a total of 2,353 annotations.

Table 2 illustrates the relative SDQC distribution for the whole dataset for both response types, being targeted towards respectively submission

Target \ Label	S	D	Q	C
Submission post	273	300	81	2,353
Parent comment	261	632	304	1,810
Submission post %	9.1	10	2.7	78.2
Parent comment %	8.7	21	10.1	60.2

Table 2: SDQC stance label distribution in DAST

(source) and parent posts, i.e. the posts replied to. The distribution is quite skewed towards the "commenting" class label, with a total distribution of S(0.091), D(0.1), Q(0.027) and C(0.782).

Rumour data: The dataset contains 16 rumourous submissions, 3 of which are true, 3 are false and the remaining 10 are unverified. They make up 220 Reddit conversations, or 596 branches, with a total of 1,489 posts, equal to about half of the dataset. The posts are distributed across the nine events as follows: 5G (233), Donald Trump (140), ISIS (169), "Kost" (324), MeToo (60), Peter Madsen (381), "Politik" (49), "Togstrejke" (73), and "Ulve i DK" (56). Thus ISIS, MeToo, and Peter Madsen are the only events which only contain rumourous conversations.

Although rumours with known truth value would be optimal for veracity classification, this might reflect reality as the truth value of rumours may stay unverified. The amount of unverified rumours does however warrant more investigation in order to use all of the rumourous submissions for rumour veracity classification. Further details about the approach to unverified rumours are covered in Section 4.

In total the dataset contains 3,007 Reddit posts distributed across 33 submissions respectively grouped into 16 events.

The tools[9] and annotated corpora (Lillie and Middelboe, 2019a) are openly released with this paper in GDPR-compliant, non-identifying format. See appendix for data statement (Bender and Friedman, 2018).

4 Method

Our veracity prediction approach depends on two components: a stance classifier and a veracity classification component (Zubiaga et al., 2018).

4.1 Stance Classification

For stance classification two different approaches have been used, one being an LSTM classifier inspired by (Kochkina et al., 2017) and the other employing a number of classic machine learning models with a focus on feature engineering as presented in Aker et al. (2017).

LSTM classifier: The LSTM model is widely used for tasks where the sequence of data and earlier elements in sequences are important (Goldberg, 2016). The temporal sequence of tweets was one of the motivations for Kochkina et al. (2017) to use the LSTM model for branches of tweets, as well as for the bidirectional conditional LSTM for Augenstein et al. (2016).

While the results from both the Bi-LSTM in Augenstein et al. (2016) and Branch-LSTM in Kochkina et al. (2017) achieve state-of-the-art performance, they both note that their deep learning approaches suffer from the lack of a larger training dataset. This is not uncommon in this task (Taulé et al., 2017; Zubiaga et al., 2016; Gorrell et al., 2019). We suspect that we would observe the same tendency for the DAST dataset, which is relatively small with its 3,007 Reddit posts. However, as the LSTM approach still manages to achieve state-of-the-art performance, we opted to include an LSTM implementation for the stance classification task.

Specifically, the LSTM classifier used for stance classification here is implemented with PyTorch[10] and consists of a number of LSTM layers and a number of ReLU layers, followed by a dropout layer and a softmax layer to perform classifications. The model is trained with stochastic gradient descent (SGD) and a negative log likelihood loss function. The configurations considered and overall approach is inspired by the Branch-LSTM classifier in (Kochkina et al., 2017), except that we do not input data grouped sequentially by branches, but one by one.

Non-neural network classifiers: It is the intention to use non-neural network models in contrast to the LSTM deep learning approach above, as research shows that this approach can do very well (Derczynski et al., 2017), and particularly Decision Tree and Random Forest classifiers (Aker et al., 2017). Furthermore Support Vector Machine (SVM) and Logistic Regression have proven to be efficient (Enayet and El-Beltagy, 2017; Derczynski et al., 2017). The models are listed below, prefixed with a label, which we will use to denote them throughout the paper:

[9] github.com/danish-stance-detectors

[10] https://pytorch.org/

logit *Logistic Regression* classifier

tree *Decision Tree* classifier

svm *Support Vector Machine* (linear kernel)

rf *Random Forest* classifier

Baselines: A simple majority voter (**MV**) as well as a stratified classifier (**SC**) were implemented. The former predicts only the most frequent class, and the latter generates predictions by respecting the training set's class distribution.

The non-neural networks models and baseline models described above are all implemented with the Scikit Learn (Pedregosa et al., 2011) framework, which provides a wide variety of machine learning implementations.

Preprocessing: As a preprocessing step, all post texts are lower-cased and then tokenised with the NLTK library (Bird et al., 2009), and finally all punctuation is removed, not including cases such as commas and periods in numbers, as well as periods in abbreviations. Furthermore URLs are replaced with the tag "urlurlurl" and quotes with the tag "refrefref".

Features: In order to represent the features of the preprocessed data numerically we employ eight feature categories, which are grouped by how they relate: text, lexicon, sentiment, Reddit, most frequent words, BoW, POS, and word embeddings. Note that only the Reddit specific features are domain-dependent, while the others should apply for the general case. The choices of features are a compilation of select features from various state-of-the-art systems (Aker et al., 2017; Kochkina et al., 2017; Enayet and El-Beltagy, 2017), except for the Reddit specific ones. Most of the features are binary, taking either a 0 or a 1 as value, and those that are not are min-max normalised (Han et al., 2011, p. 114), except for the word embeddings.

Table 3 presents an overview of the total feature vector, including the feature categories and their number of individual features. Note that the word embeddings are actually 300 long, but the extra 3 features are the cosine similarities between different word embeddings with regards to parent, source, and branch word tokens.

Sentiment analysis is performed with the Afinn library (Årup Nielsen, 2011), and POS tagging is performed with the Danish Polyglot library (Al-Rfou et al., 2013). Text features include binary features for presence of: '.', '!', '?', 'hv'-words, '...', as well as text length, URL count, maximum length of capital character sequence, and count of: '...', '?', '!', and words. Finally the text features include ratio of capital letters to non-capital letters and average word length.

Category	*Length*
Text	13
Lexicon	4
Sentiment	1
Reddit	10
Most frequent words	132
BOW	13,663
POS	17
Word embeddings	303
Total	14,143

Table 3: Feature vector overview

Lexicon features are extracted by looking up occurrences of items in four predefined lexicon/dictionaries: negation words, swear words, positive smileys, and negative smileys. Negation words are translated from the English list used in Kochkina et al. (2017), as no list could be found for this purpose elsewhere. Beyond ourselves, swear words are taken from various sources: youswear.com, livsstil.tv2.dk, dansk-og-svensk.dk, and dagens.dk. Smiley lists were compiled from Wikipedia using the western style emoticons.[11]

Reddit-specific features include karma, gold status, Reddit employment status (if any), verified e-mail, reply count, upvotes, and whether the user is the submission submitter. Further, based on Reddit commenting syntax, the following features are included: sarcasm ('/s'), edited ('edit:'), and quote count ('>').

Finally, word embeddings are generated with word2vec (Mikolov et al., 2013) using the Gensim framework (Řehůřek and Sojka, 2010). The word vectors are trained on a Danish text corpus acquired from "Det Danske Sprog- og Litteraturselskab" (DSL),[12] consisting of 45 million tokens of written LGP (Language for General Purposes),[13] as well as the preprocessed Reddit text.

4.2 Rumour Veracity Prediction

The rumour veracity classification implemented is inspired by the approach presented in Dungs et al. (2018). This approach is especially interesting

[11] en.wikipedia.org/wiki/List_of_emoticons

[12] https://dsl.dk/

[13] https://korpus.dsl.dk/resources.html

as it relinquishes language specific features such as word embeddings and relies on sequences of stance labels and temporal features of the posts. This possibly enables the use of data across languages and platforms. One implemented HMM, λ is alike the model presented in Dungs et al. (2018) receiving sequences of stance labels ordered by their posting time as input. For each label presented in the data, a model is trained with training data for that label. For example in a label space containing the labels "True" and "False", two HMM models λ_{false} and λ_{true} are trained. The predicted label of the model λ will be whichever labelled model presents a higher probability score for the given sequence.

Further a model ω is built, which differs from λ by also containing normalised timestamps for each post. This was done as inclusion of temporal features boosted performance in (Dungs et al., 2018). Note that ω is not the same model as the variably-spaced λ' of Dungs et al. (2018).

As a baseline throughout the experiments, a simple stratified baseline will be used, denoted *VB*. The baseline notes the average distribution of stance labels as a four-tuple for respectively true and false (and unverified where relevant) rumours. When predicting rumour veracity, *VB* calculates the distribution of stance labels in a given sequence in the testing data and chooses the truth value with the most similar class label distribution.

The data used for experiments across languages and platforms include the PHEME dataset (Zubiaga et al., 2016; Derczynski et al., 2015). First, experiments are performed isolated on DAST. Then, the PHEME dataset is used as training data while DAST is used as test set. Further, unverified rumours are approached in two ways: (1) three-way classification is performed on true, false *and* unverified rumours, and (2) two-way classification is performed with unverified rumours treated as True. The results are presented in Section 5.2.

The data from DAST is used in three different ways, given the expected discrepancies between the English Twitter data and the Danish Reddit data. The Reddit conversation structure in Figure 1 differs slightly from the Twitter structure. The submission post is the actual source of conversation, while conversation top level comments are the source for Twitter conversations. Three different representations are tested for DAST:

BAS each branch in a conversation is regarded as a rumour (**b**ranch-**a**s-**s**ource). This causes partial duplication of comments, as branches can share parent comments.

TCAS top level comments are regarded as the source of a rumour and the conversation tree they spawn are the sequences of labels (**t**op-level **c**omment-**a**s-**s**ource).

SAS the entire submission is regarded as a rumour (**s**ubmission-**a**s-**s**ource). The SAS approach means that only 16 instances are available.

4.3 Evaluation Measures

Most of the related work report results with accuracy as scoring metric (Derczynski et al., 2017; Aker et al., 2017), which expresses the ratio of number of correct predictions to the total number of input samples. However, this becomes quite uninteresting if the input samples have imbalanced class distributions, which is the case for our dataset. What *is* interesting to measure is how well the models are at predicting the correct class labels. As such, in addition to reporting accuracy we will also use the F_1 scoring metric. In particular we will use an unweighted macro-averaged F_1 score for the case of multi-class classification.

5 Results and Analysis

This section reports performance at stance classification and rumour veracity prediction.

5.1 Stance Classification Results

First, an ablation study of the feature groups revealed that the Reddit specific features as well as lexicon features contributed negatively to performance for stance classification. Further, it turned out that the Most Frequent Words (MFW) feature category resembled BOW with low variance features removed. Finally the generated MFW list contained stopwords very specific to DAST, such as "B12", "CO2", and "5G". As such all classifiers has the feature categories Reddit, lexicon, and MFW removed.

Second, parameter search was performed through grid-search for three classifiers, being LSTM, Logistic Regression (**logit**), and Support Vector Machine (**svm**). Decision Tree and Random Forest were omitted due to poor performance. Full details of the parameters searched are given in Middelboe and Lillie (2019).

Classifier results are given in Table 4, under 5-fold (stratified) cross validation. Top-level com-

Model	Macro-F_1	σ	Accuracy	σ
MV	0.2195	(+/- 0.00)	0.7825	(+/- 0.00)
SC	0.2544	(+/- 0.04)	0.6255	(+/- 0.01)
logit	0.3778	(+/- 0.06)	0.7812	(+/- 0.02)
svm	0.3982	(+/- 0.04)	0.7496	(+/- 0.02)
logit'	0.4112	(+/- 0.07)	0.7549	(+/- 0.04)
svm'	**0.4212**	(+/- 0.06)	0.7572	(+/- 0.02)
LSTM	0.2802	(+/- 0.04)	0.7605	(+/- 0.03)
LSTM'	0.3060	(+/- 0.05)	0.7163	(+/- 0.16)

Table 4: Stance cross validation results for *logit*, *svm*, LSTM, and baselines with macro F_1 and accuracy, including standard deviation (σ).

Structure	Model	Acc.	F_1
SAS	λ	**0.81**	**0.45**
	ω	0.81	0.45
	VB	0.39	0.36
TCAS	λ	0.73	**0.63**
	ω	**0.79**	0.61
	VB	0.35	0.35
BAS	λ	0.78	0.66
	ω	**0.83**	**0.68**
	VB	0.43	0.42

Table 5: Stance-only veracity prediction, cross-validated over the Danish-language DAST corpus.

ments are not shared across splits. The baselines **MV** and **SC** and the default models for Logistic Regression (*logit*) and Support Vector Machine (*svm*) are included. *logit*' and *svm*' denote parameter-tuned models without Reddit, lexicon, and MFW features. Finally *LSTM* is the parameter-tuned model with all features; *LSTM*' is without Reddit, lexicon, or MFW features.

We see that *svm*' is the best performing model, achieving a macro F_1 score of 0.42, an improvement of 0.02 over the default model. Note that the accuracy is worse than the *MV* baseline, and *logit*' has even decreased its accuracy. The reason for this could be that the models have been tuned specifically for macro F_1. As expected we see that *MV* only predicts "commenting" classes and that *SC* follows the class label distribution of the dataset, while *logit*' and *svm*' are able to predict the under-represented classes. Because of the low-volume of data in DAST we did not expect the LSTM to perform very well, which was reflected in its best macro F_1 score of 0.3060.

5.2 Rumour Veracity Prediction Results

The results for the rumour veracity experiments are presented in this section. For size limitation reasons only the result tables for "Unverified" rumours interpreted as "False" are included, as these are superior. When interpreting "Unverified" as "True" the overall results for the experiments are worse. This indicates that the stance sequences in "Unverified" and "False" rumours are more alike than those in "True" rumours. When performing 3-fold cross validation on DAST, the best results are observed with the ω model on the BAS structure with an accuracy of 0.83 and an F_1 of 0.68.

We hypothesise that stance structures leading to veracity predictions may be similar across languages. To investigate this, we trained HMMs using the PHEME data (mostly English and German) and evaluated performance of these models over DAST. Results are in Table 6.

Structure	Model	Acc.	F_1
SAS	λ	**0.88**	**0.71**
	ω	0.75	0.67
	VB	0.81	0.45
TCAS	λ	0.77	**0.66**
	ω	**0.81**	0.59
	VB	0.80	0.62
BAS	λ	**0.82**	**0.67**
	ω	0.67	0.57
	VB	0.77	0.53

Table 6: Veracity prediction from stance only, training on English/German PHEME rumour discussions and testing on Danish-language DAST.

The best performance is under the SAS structure. Note that the results when transferring veracity prediction across languages not only match the in-language training, but in fact exceed it. This indicates that cross-lingual stance transfer is possible with advantages, suggesting extra-lingual behaviours present in conversations about true and false claims. The increase in performance is attributed to the larger amount of training size available in the PHEME dataset compared to performing cross-validation within DAST.

There is also an interesting note about the effect of post times. λ performs better than ω when training on PHEME data, but ω performs better when solely training and testing on DAST. This suggests differences in the posting time tendencies, which may be caused by either the platform or language differences between the datasets.

5.3 Joining Stance and Veracity Prediction

To investigate the use of the system on new unseen data, the SVM stance classifier is used to classify stance labels for each of the rumour submissions. This is done by training on all of DAST except the one rumour to classify stance for ("hold one out").

Structure	Model	Acc.	F_1
SAS	λ	0.81	0.64
	ω	0.75	**0.67**
	VB	0.81	0.45
TCAS	λ	0.79	0.56
	ω	0.68	0.55
	VB	0.76	0.43
BAS	λ	**0.82**	0.58
	ω	0.76	0.56
	VB	0.76	0.48

Table 7: Training on the PHEME dataset and testing on automatic stance labels generated for DAST with "Unverified" rumours treated as "False".

This use of predicted instead of gold stance labels evaluates the system's extrinsic performance.

The best results were seen with "Unverified" labels as false, with the λ model on the BAS structure, which is reported in Table 7. A general tendency compared to the gold label results is a marginal drop in F_1, but little to no effect in the veracity prediction performance of the system.

5.4 Unverified as True

In the following experiments the unverified rumours have been interpreted as true rumours. Comparisons between these results and the 'Unverified as false' experiments above, might reveal interesting properties about the data. Switching between interpreting unverified as true or as false should approximately afford either higher rumour detection precision or higher recall, respectively.

Structure	Model	Acc.	F_1
SAS	λ	0.74 (+/- 0.21)	0.49 (+/- 0.13)
	ω	**0.74** (+/- 0.21)	0.53 (+/- 0.33)
	VB	0.19 (+/- 0.03)	0.16 (+/- 0.02)
TCAS	λ	0.67 (+/- 0.09)	0.55 (+/- 0.08)
	ω	0.65 (+/- 0.16)	0.49 (+/- 0.16)
	VB	0.34 (+/- 0.02)	0.34 (+/- 0.02)
BAS	λ	0.61 (+/- 0.05)	0.54 (+/- 0.07)
	ω	0.71 (+/- 0.06)	**0.62** (+/- 0.05)
	VB	0.59 (+/- 0.10)	0.54 (+/- 0.03)

Table 8: Danish veracity results on 3-fold cross validation for unverified being true.

Results are given in Table 8. This framing generally saw lower scores than comparable prior results (i.e. Table 5), with the highest accuracy at 0.74 achieved with the ω and λ models on the SAS structure. The highest F_1 score is achieved by ω on BAS, at 0.62.

To check if this result was specific to Danish, we repeated the experiment, over the English and German conversations in the larger PHEME dataset, again using its gold stance labels. Results are in Table 9. The performance level held in this non-Danish setting. The highest accuracy achieved is 0.81 reached by the ω model on the SAS structure. The highest F_1 score is 0.59, achieved on the SAS structure as well by the λ model.

Structure	Model	Acc.	F_1
SAS	λ	0.75	**0.59**
	ω	**0.81**	0.45
	VB	0.69	0.54
TCAS	λ	0.72	0.54
	ω	0.76	0.52
	VB	0.70	0.56
BAS	λ	0.62	0.56
	ω	0.60	0.51
	VB	0.61	0.58

Table 9: Training and testing on PHEME data, with "Unverified" rumours treated as "True".

5.5 Usage Implications

The consequences of declaring a claim to be true or false can be serious. As in Derczynski et al. (2019), we intend this technology to be used solely as part of a "human-in-the-loop" system; although stories may be flagged automatically as false (or true), these should be presented to humans as unreliable results for analysis. On the other hand, technology offers potential to assist in the vital task of finding candidate misinformation among vast amounts of web data.

6 Conclusion

Social media has created a platform for the spread of rumours, which are stories with unverified claims. We investigated how to automatically predict the veracity of rumours spread on Danish social media by analysing the stance of conversation participants. Through experiments a Linear SVM gave SDQC stance classification with an accuracy of 0.76 and a macro F_1 of 0.42. An HMM then predicted rumour veracity automatically-labelled stance with up to 81% accuracy.

Interestingly, we find that veracity prediction models that use only stance labels from conversations in one language can be transferred effectively to predict veracity in conversations held in another language, based again on stance. This indicates the presence and utility of cross-lingual conversational behaviours around true and false claims.

Further and extensive experimentation and results can be found in the thesis that led to this work (Middelboe and Lillie, 2019).

References

Ahmet Aker, Leon Derczynski, and Kalina Bontcheva. 2017. Simple Open Stance Classification for Rumour Analysis. In *Proceedings of Recent Advances in Natural Language Processing*, pages 31–39, Varna, Bulgaria.

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 183–192, Sofia, Bulgaria. Association for Computational Linguistics.

Pepa Atanasova, Preslav Nakov, Lluís Màrquez, Alberto Barrón-Cedeño, Georgi Karadzhov, Tsvetomila Mihaylova, Mitra Mohtarami, and James Glass. 2019. Automatic fact-checking using context and discourse information. *Journal of Data and Information Quality*, 11(3):12.

Isabelle Augenstein, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. Stance Detection with Bidirectional Conditional Encoding. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 876–885, Austin, Texas.

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Steven Bird, Edward Loper, and Ewan Klein. 2009. Natural language processing with python.

Leon Derczynski, Torben Oskar Albert-Lindqvist, Marius Venø Bendsen, Nanna Inie, Viktor Due Pedersen, and Jens Egholm Pedersen. 2019. Misinformation on twitter during the danish national election. In *Proceedings of the conference for Truth and Trust Online*.

Leon Derczynski, Kalina Bontcheva, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Arkaitz Zubiaga. 2017. SemEval-2017 Task 8: RumourEval: Determining rumour veracity and support for rumours. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 69–76, Vancouver, Canada.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2015. PHEME: Computing Veracity—the Fourth Challenge of Big Social Data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session*.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can Rumour Stance Alone Predict Veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, page 3360–3370, Santa Fe, New Mexico, USA.

Omar Enayet and Samhaa R. El-Beltagy. 2017. NileTMRG at SemEval-2017 Task 8: Determining Rumour and Veracity Support for Rumours on Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 470–474, Vancouver, Canada.

Martin Fajcik, Pavel Smrz, and Lukas Burget. 2019. BUT-FIT at SemEval-2019 Task 7: Determining the Rumour Stance with Pre-Trained Deep Bidirectional Transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*.

Yoav Goldberg. 2016. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research 57*, pages 345–420. Chapters 10-11.

Genevieve Gorrell, Elena Kochkina, Maria Liakata, Ahmet Aker, Arkaitz Zubiaga, Kalina Bontcheva, and Leon Derczynski. 2019. SemEval-2019 task 7: RumourEval, determining rumour veracity and support for rumours. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 845–854. Association for Computational Linguistics.

Jiawei Han, Micheline Kamber, and Jian Pei. 2011. *Data Mining, Concepts and Techniques*, 3 edition. Morgan Kaufmann Publishers Inc.

Y. Linlin Huang, Kate Starbird, Mania Orand, Stephanie A. Stanek, and Heather T. Pedersen. 2015. Connected through crisis: Emotional proximity and the spread of misinformation online. In *CSCW '15 Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing*, pages 969–980, Vancouver, BC, Canada. Association for Computing Machinery.

Andreas Kirkedal, Barbara Plank, Leon Derczynski, and Natalie Schluter. 2019. The Lacunae of Danish Natural Language Processing. In *Proceedings of the Nordic Conference on Computational Linguistics (NODALIDA)*. Northern European Association for Language Technology.

Elena Kochkina, Maria Liakata, and Isabelle Augenstein. 2017. Turing at SemEval-2017 Task 8: Sequential Approach to Rumour Stance Classification with Branch-LSTM. In *Proceedings of the 11th International Workshop on Semantic Evaluations (SemEval-2017)*, pages 475–480, Vancouver, Canada.

Rasmus Lehmann and Leon Derczynski. 2019. Political Stance in Danish. In *Proceedings of the Nordic Conference on Computational Linguistics (NODALIDA)*. Northern European Association for Language Technology.

Anders Edelbo Lillie and Emil Refsgaard Middelboe. 2019a. Danish stance-annotated Reddit dataset. doi: 10.6084/m9.figshare.8217137.v1.

Anders Edelbo Lillie and Emil Refsgaard Middelboe. 2019b. Fake news detection using stance classification: A survey. *arXiv preprint arXiv:1907.00181*.

Emil Refsgaard Middelboe and Anders Edelbo Lillie. 2019. Danish Stance Classification and Rumour Resolution. Master's thesis, ITU Copenhagen.

Tomas Mikolov, G.s Corrado, Kai Chen, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *Proceedings of the International Conference on Learning Representations (ICLR 2013)*, pages 1–12. arXiv preprint arXiv:1301.3781.

Saif M. Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of SemEval-2016*, pages 31–41, San Diego, California.

Finn Årup Nielsen. 2011. A new ANEW: evaluation of a word list for sentiment analysis in microblogs. In *Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big things come in small packages. Volume 718 in CEUR Workshop Proceedings*, pages 93–98.

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

Vahed Qazvinian, Emily Rosengren, Dragomir R. Radev, and Qiaozhu Mei. 2011. Rumor has it: Identifying Misinformation in Microblogs. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1589–1599, Edinburgh, Scotland.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. `http://is.muni.cz/publication/884893/en`.

Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. 2017. Fake news detection on social media: A data mining perspective. *ACM SIGKDD Explorations Newsletter*, 19(1):22–36.

Mariona Taulé, M. Antónia Martí, Francisco Rangel, Paolo Rosso, Cristina Bosco, and Viviana Patti. 2017. Overview of the Task on Stance and Gender Detection in Tweets on Catalan Independence at IberEval 2017. In *Proceedings of the Second Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2017)*, pages 157–177.

James Thorne, Andreas Vlachos, Oana Cocarascu, Christos Christodoulopoulos, and Arpit Mittal. 2018. The Fact Extraction and VERification (FEVER) shared task. In *Proceedings of the First Workshop on Fact Extraction and VERification*.

Brian Xu, Mitra Mohtarami, and James Glass. 2019. Adversarial domain adaptation for stance detection. In *Proceedings of the Conference on Neural Information Processing Systems*, Montréal, Canada.

Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. 2018. Detection and Resolution of Rumours in Social Media: A Survey. *ACM Computing Surveys, Vol. 51, No. 2, Article 32*.

Arkeitz Zubiaga, Maria Liakata, Rob Procter, Geraldine Wong Sak Hoi, and Peter Tolmie. 2016. Analysing How People Orient to and Spread Rumours in Social Media by Looking at Conversational Threads. *PLoS ONE*. 11(3).

A Rumour sources

Event	*Submission title*	*Rumour status*
5G	5G-teknologien er en miljøtrussel, som bør stoppes	Unverified
	Det er ikke alle, som glæder sig til 5G.	Unverified
	Uffe Elbæk er bekymret over de "sundhedsmæssige konsekvenser" af 5G-netværket	Unverified
Donald Trump	Hvorfor må DR skrive sådan noget åbenlyst falsk propaganda?	Unverified
	16-årig blev anholdt for at råbe 'fuck Trump' til lovlig demonstration mod Trump	Unverified
ISIS	23-årig dansk pige har en dusør på $1 million på hendes hovede efter at have dræbt mange ISIS militanter	Unverified
	Danish student 'who killed 100 ISIS militants has $1million bounty on her head but is treated as terrorist' (The Mirror)	Unverified
Kost	Bjørn Lomborg: Du kan være vegetar af mange gode grunde - men klimaet er ikke en af dem	Unverified
	Professor: Vegansk kost kan skade småbørns vækst	False
MeToo	Björks FB post om Lars Von Trier (#MeToo)	Unverified
Peter Madsen	Savnet ubåd er fundet i Køge Bugt: Alle er i god behold	False
	Undersøgelser bekræfter: Ubåd blev angiveligt sunket bevidst	True
	Peter Madsen: Kim Wall døde i en ulykke på ubåden	False
Politik	KORRUPT	True
Togstrejke	De ansatte i DSB melder om arbejdsnedlæggelse 1. april.	True
Ulve i DK	Den vedholdende konspirationsteori: Har nogen udsat ulve i Nordjylland?	Unverified

Table 10: Overview of the rumour submissions and their veracity.

B Veracity results

Structure	Model	Acc.	F_1
SAS	λ	0.53 (+/- 0.09)	0.53 (+/- 0.10)
	ω	0.55 (+/- 0.09)	0.55 (+/- 0.10)
	VB	0.37 (+/- 0.03)	0.31 (+/- 0.07)
TCAS	λ	0.60 (+/- 0.07)	0.58 (+/- 0.08)
	ω	0.64 (+/- 0.05)	0.61 (+/- 0.05)
	VB	0.53 (+/- 0.04)	0.38 (+/- 0.03)
BAS	λ	0.60 (+/- 0.05)	0.58 (+/- 0.05)
	ω	**0.67** (+/- 0.03)	**0.62** (+/- 0.04)
	VB	0.49 (+/- 0.10)	0.40 (+/- 0.01)
None	λ	0.55 (+/- 0.05)	0.54 (+/- 0.07)
	ω	0.57 (+/- 0.08)	0.55 (+/- 0.10)
	VB	0.43 (+/- 0.03)	0.33 (+/- 0.08)

Table 11: Training and testing on mix of PHEME data and different DAST structures for unverified false

C Data statement

Curation rationale Comments around rumourous claims.

Language variety BCP-47: `da-DK`

Speaker demographic

- Reddit users
- Age: Unknown – mixed.
- Gender: Unknown – mixed.
- Race/ethnicity: Unknown – mixed.
- Native language: Unknown; Danish speakers.
- Socioeconomic status: Unknown – mixed.
- Different speakers represented: Unknown; upper bound is the number of posts.
- Presence of disordered speech: Rare.

Annotator demographic

- Age: 20-30.
- Gender: male.
- Race/ethnicity: white northern European.
- Native language: Danish.
- Socioeconomic status: higher education student.

Speech situation Discussions held in public on the Reddit platform.

Text characteristics Danish colloquial web speech.

Provenance Originally taken from Reddit, 2018.