# TDParse: Multi-target-specific sentiment recognition on Twitter

**Bo Wang** **Maria Liakata** **Arkaitz Zubiaga** **Rob Procter**
Department of Computer Science
University of Warwick
Coventry, UK
{bo.wang, m.liakata, a.zubiaga}@warwick.ac.uk

## Abstract

Existing target-specific sentiment recognition methods consider only a single target per tweet, and have been shown to miss nearly half of the actual targets mentioned. We present a corpus of UK election tweets, with an average of 3.09 entities per tweet and more than one type of sentiment in half of the tweets. This requires a method for multi-target specific sentiment recognition, which we develop by using the context around a target as well as syntactic dependencies involving the target. We present results of our method on both a benchmark corpus of single targets and the multi-target election corpus, showing state-of-the art performance in both corpora and outperforming previous approaches to multi-target sentiment task as well as deep learning models for single-target sentiment.

## 1 Introduction

Recent years have seen increasing interest in mining Twitter to assess public opinion on political affairs and controversial issues (Tumasjan et al., May 2010; Wang et al., 2012) as well as products and brands (Pak and Paroubek, 2010). Opinion mining from Twitter is usually achieved by determining the overall sentiment expressed in an entire tweet. However, inferring the sentiment towards specific targets (e.g. people or organisations) is severely limited by such an approach since a tweet may contain different types of sentiment expressed towards each of the targets mentioned. An early study by Jiang et al. (2011) showed that 40% of classification errors are caused by using tweet-level approaches that are independent of the target. Consider the tweet:

> *"I will b voting 4* ***Greens*** *... 1st reason: 2 remove 2 party alt. of* ***labour*** *or* ***conservative*** *every 5 years. 2nd:* ***fracking****"*

The overall sentiment is positive but there is a negative sentiment towards "labour", "conservative" and "fracking" and a positive sentiment towards "Greens". Examples like this are common in tweets discussing topics like politics. As has been demonstrated by the failure of election polls in both referenda and general elections (Burnap et al., 2016), it is important to understand not only the overall mood of the electorate, but also to distinguish and identify sentiment towards different key issues and entities, many of which are discussed on social media on the run up to elections.

Recent developments on target-specific Twitter sentiment classification have explored different ways of modelling the association between target entities and their contexts. Jiang et al. (2011) propose a rule-based approach that utilises dependency parsing and contextual tweets. Dong et al. (2014), Tang et al. (2016a) and Zhang et al. (2016) have studied the use of different recurrent neural network models for such a task but the gain in performance from the complex neural architectures is rather unclear[1]

In this work we introduce the multi-target-specific sentiment recognition task, building a corpus of tweets from the 2015 UK general election campaign suited to the task. In this dataset, target entities have been semi-automatically selected, and sentiment expressed towards multiple target entities as well as high-level topics in a tweet have been manually annotated. Unlike all existing studies on target-specific Twitter sentiment analysis, we move away from the assumption that

[1]They have yet to show a clear out-performance on a benchmarking dataset and our multi-target corpus, possibly because they usually require large amount of training data.



*Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 483–493,
Valencia, Spain, April 3-7, 2017. 

each tweet mentions a single target; we introduce a more realistic and challenging task of identifying sentiment towards multiple targets within a tweet. To tackle this task, we propose TDParse, a method that divides a tweet into different segments building on the approach introduced by Vo and Zhang (2015). TDParse exploits a syntactic dependency parser designed explicitly for tweets (Kong et al., 2014), and combines syntactic information for each target with its left-right context.

We evaluate and compare our proposed system both on our new multi-target UK election dataset, as well as on the benchmarking dataset for single-target dependent sentiment (Dong et al., 2014). We show a clear state-of-the-art performance of TDParse over existing approaches for tweets with multiple targets, which encourages further research on the multi-target-specific sentiment recognition task.[2]

## 2 Related Work: Target-dependent Sentiment Classification on Twitter

The 2015 Semeval challenge introduced a task on target-specific Twitter sentiment (Rosenthal et al., 2015) which most systems (Boag et al., 2015; Plotnikova et al., 2015) treated in the same way as tweet level sentiment. The best performing system in the 2016 Semeval Twitter challenge substask B (Nakov et al., 2016), named Tweester, also performs on tweet level sentiment classification. This is unsurprising since tweets in both tasks only contain a single predefined target entity and as a result often a tweet-level approach is sufficient. An exception to tweet level approaches for this task, showing promise, is Townsend et al. (2015), who trained a SVM classifier for tweet segmentation, then used a phrase-based sentiment classifier for assigning sentiment around the target. The Semeval aspect-based sentiment analysis task (Pontiki et al., 2015; Pateria and Choubey, 2016) aims to identify sentiment towards entity-attribute pairs in customer reviews. This differs from our goal in the following way: both the entities and attributes are limited to a predefined inventory of limited size; they are aspect categories reflected in the reviews rather than specific targets, while each review only has one target entity, e.g. a laptop or a restaurant. Also sentiment classification in formal text such as product reviews is very different from that in tweets. Recently Vargas et al. (2016) analysed the differences between the overall and target-dependent sentiment of tweets for three events containing 30 targets, showing many significant differences between the corresponding overall and target-dependent sentiment labels, thus confirming that these are distinct tasks.

Early work tackling target-dependent sentiment in tweets (Jiang et al., 2011) designed target-dependent features manually, relying on the syntactic parse tree and a set of grammar-based rules, and incorporating the sentiment labels of related tweets to improve the classification performance. Recent work (Dong et al., 2014) used recursive neural networks and adaptively chose composition functions to combine child feature vectors according to their dependency type, to reflect sentiment signal propagation to the target. Their data-driven composition selection approach replies on the dependency types as features and a small set of rules for constructing target-dependent trees. Their manually annotated dataset contains only one target per tweet and has since been used for benchmarking by several subsequent studies (Vo and Zhang, 2015; Tang et al., 2016a; Zhang et al., 2016). Vo and Zhang (2015) exploit the left and right context around a target in a tweet and combine low-dimensional embedding features from both contexts and the full tweet using a number of different pooling functions. Despite not fully capturing semantic and syntactic information given the target entity, they show a much better performance than Dong et al. (2014), indicating useful signals in relation to the target can be drawn from such context representation. Both Tang et al. (2016a) and Zhang et al. (2016) adopt and integrate left-right target-dependent context into their recurrent neural network (RNN) respectively. While Tang et al (2016a) propose two long short-term memory (LSTM) models showing competitive performance to Vo and Zhang (2015), Zhang et al (2016) design a gated neural network layer between the left and right context in a deep neural network structure but require a combination of three corpora for training and evaluation. Results show that conventional neural network models like LSTM are incapable of explicitly capturing important context information of a target (Tang et al., 2016b). Tang et al. (2016a) also experiment with adding attention layers for LSTM but

[2]The data and code can be found at `https://goo.gl/S2T1GO`

fail to achieve competitive results possibly due to the small training corpus.

Going beyond the existing work we study the more challenging task of classifying sentiment towards multiple target entities within a tweet. Using the syntactic information drawn from tweet-specific parsing, in conjunction with the left-right contexts, we show the state-of-the-art performance in both single and multi-target classification tasks. We also show that the tweet level approach that many sentiment systems adopted in both Semeval challenges, fail to capture all target-sentiments in a multi-target scenario (Section 5.1).

## 3 Creating a Corpus for Target Specific Sentiment in Twitter

We describe the design, collection and annotation of a corpus of tweets about the 2015 UK election.

### 3.1 Data Harvesting and Entity Recognition

We collected a corpus of tweets about the UK elections, as we wanted to select a political event that would trigger discussions on multiple entities and topics. Collection was performed through Twitter's streaming API and tracking 14 hashtags[3]. Data harvesting was performed between 7th February and 30th March 2015. This led to the collection of 712k tweets, from which a subset was sampled for manual annotation of target-specific sentiment. We also created a list of 438 topic keywords relevant to 9 popular election issues[4] for data sampling. The initial list of 438 seed words provided by a team of journalists was augmented by searching for similar words within a vector space on the basis of cosine similarity. Keywords are used both in order to identify thematically relevant tweets and also targets. We also consider named entities as targets.

Sampling of tweets was performed by removing retweets and making sure each tweet contained at least one topic keyword from one of the 9 election issues, leading to 52,190 highly relevant tweets. For the latter we ranked tweets based on a "similarity" relation, where "similarity" is measured as a function of content overlap (Mihalcea, 2004). Formally, given a tweet $S_i$ being represented by the set of $N$ words that appear in the tweet: $S_i = W_i^1, W_i^2, ..., W_i^N$ and our list of curated topic keywords $T$, the ranking function is defined as:

$$log(|S_i|) * |W_i \in S_i \cap W_i \in T| \quad (1)$$

where $|S_i|$ is the total number of words in the tweet; unlike Mihalcea (2004) we prefer longer tweets. We used exact matching with flexibility on the special characters at either end. TF-IDF normalisation and cosine similarity were then applied to the dataset to remove very similar tweets (empirically we set the cosine similarity threshold to 0.6). We also collected all external URLs mentioned in our dataset and their web content throughout the data harvesting period, filtering out tweets that only contain an external link or snippets of a web page. Finally we sampled 4,500 top-ranked tweets keeping the representation of tweets mentioning each election issue proportionate to the original dataset.

For annotation we considered sentiment towards two types of targets: entities and topic keywords. Entities were processed in two ways: firstly, named entities (people, locations, and organisations) were automatically annotated by combining the output of Stanford Named Entity Recognition (NER) (Finkel et al., 2005), NLTK NER (Bird, 2006) and a Twitter-specific NER (Ritter et al., 2011). All three were combined for a more complete coverage of entities mentioned in tweets and subsequently corrected by removing wrongly marked entities through manual annotation. Secondly, to make sure we covered all key entities in the tweets, we also matched tweets against a manually curated list of 7 political-party names and added users mentioned therein as entities. The second type of targets matched the topic keywords from our curated list.

### 3.2 Manual Annotation of Target Specific Sentiment

We developed a tool for manual annotation of sentiment towards the targets (i.e. entities and topic keywords) mentioned in each tweet. The annotation was performed by nine PhD-level journalism students, each of them annotating approximately a ninth of the dataset, i.e. 500 tweets. Additionally, they annotated a common subset of 500 tweets consistign of 2,197 target entities, which was used to measure inter-annotator agreement (IAA). An-

[3]#ukelection2015, #ge2015, #ukge2015, #ukgeneralelection2015, #bbcqt, #bbcsp, #bbcdp, #marrshow, #generalelection2015, #ge15, #generalelection, #electionuk, #ukelection and #electionuk2015

[4]EU and immigration, economy, NHS, education, crime, housing, defense, public spending, environment and energy

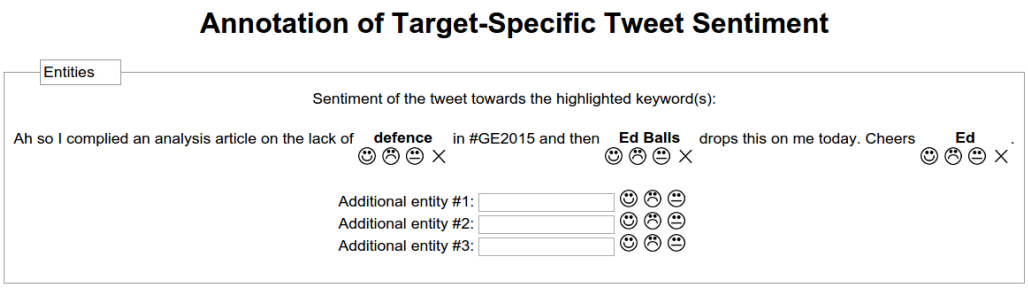


Figure 1: Annotation tool for human annotation of target specific sentiment analysis

notators were shown detailed guidelines[5] before taking up the task, after which they were redirected to the annotation tool itself (see Figure 1).

Tweets were shown to annotators one by one, and they had to complete the annotation of all targets in a tweet to proceed. The tool shows a tweet with the targets highlighted in bold. Possible annotation actions consisted in: (1) marking the sentiment for a target as being positive, negative, or neutral, (2) marking a target as being mistakenly highlighted (i.e. 'doesnotapply') and hence removing it, and (3) highlighting new targets that our preprocessing step had missed, and associating a sentiment value with them. In this way we obtained a corrected list of targets for each tweet, each with an associated sentiment value.

We measure inter-annotator agreement in two different ways. On the one hand, annotators achieved $\kappa = 0.345$ ($z = 92.2, p < 0.0001$) (fair agreement)[6] when choosing targets to be added or removed. On the other hand, they achieved a similar score of $\kappa = 0.341$ ($z = 77.7, p < 0.0001$) (fair agreement) when annotating the sentiment of the resulting targets. It is worth noting that the sentiment annotation for each target also involves choosing among not only positive/negative/neutral but also a fourth category 'doesnotapply'. The resulting dataset contains 4,077 tweets, with an average of 3.09 entity mentions (targets) per tweet. As many as 3,713 tweets have more than a single entity mention (target) per tweet, which makes the task different from 2015 Semeval 10 subtask C (Rosenthal et al., 2015) and a target-dependent benchmarking dataset of Dong et al. (2014) where each tweet has only one target annotated and thus one sentiment label assigned. The number of targets in the 4,077 tweets to be annotated originally amounted to 12,874. However, the annotators unhighlighted 975 of them, and added 688 new ones, so that the final number of targets in the dataset is 12,587. These are distributed as follows: 1,865 are positive, 4,707 are neutral, and 6,015 are negative. This distribution shows the tendency of a theme like politics, where users tend to have more negative opinions. This is different from the Semeval dataset, which has a majority of neutral sentiment. Looking at the annotations provided for different targets within each tweet, we observe that 2,051 tweets (50.3%) have all their targets consistently annotated with a single sentiment value, 1,753 tweets (43.0%) have two different sentiments, and 273 tweets (6.7%) have three different sentiment values. These statistics suggest that providing a single sentiment for the entire tweet would not be appropriate in nearly half of the cases confirming earlier observations (Jiang et al., 2011).

We also labelled each tweet containing one or more topics from the 9 election issues, and asked the annotators to mark the author's sentiment towards the topic. Unlike entities, topics may not be directly present in tweets. We compare topic sentiment with target/entity sentiment for 3963 tweets from our dataset adopting the approach by Vargas et al. (2016). Table 1 reports the individual $c(s_{target})$, $c(s_{topic})$ and joint $c(s_{target}, s_{topic})$ distributions of the target/entity $s_{target}$ and topic $s_{topic}$ sentiment. While $s_{target}$ and $s_{topic}$ report how often each sentiment category occurs in the dataset, the joint distribution $c(s_{target}, s_{topic})$ (the inner portions of the table) shows the discrepancies between target and topic sentiments. We observe marked differences between the two sentiment labels. For example it shows the topic sentiment is more neutral (1438.7 vs. 1104.1) and less negative (1930.7 vs. 2285.5) than the target sen-

[5]This guidelines can be found along with our released corpus: https://goo.gl/CjuHzd

[6]We report the strength of agreement using the benchmarks by Landis and Koch (1977) for interpreting Fleiss' kappa.

timent. There is also a number of tweets expressing neutrality towards the topics mentioned but polarised sentiment towards targets (i.e. we observe $c(s_{topic} = neu \cap s_{targets} = neg) = 258.6$ also $c(s_{topic} = neu \cap s_{targets} = pos) = 101.4$), and vice versa. This emphasises the importance of distinguishing target entity sentiment not only on the basis of overall tweet sentiment but also in terms of sentiment towards a topic.

| $c(s_{target}, s_{topic})$ | | $s_{topic}$ negative | neutral | positive | $c(s_{topic})$ |
|---|---|---|---|---|---|
| $s_{target}$ | negative | **1553.9** | 258.6 | 118.3 | 1930.9 |
| | neutral | 557.6 | **744.1** | 137.0 | 1438.7 |
| | positive | 174.0 | 101.4 | **318.1** | 593.5 |
| | $c(s_{target})$ | 2285.5 | 1104.1 | 573.4 | 3963.0 |

Table 1: Individual $c(s_{target})$, $c(s_{topic})$ and joint $c(s_{target}, s_{topic})$ distributions of sentiments

## 4 Developing a state-of-the-art approach for target-specific sentiment

### 4.1 Model development for single-target benchmarking data

Firstly we adopt the context-based approach by Vo and Zhang (2015), which divides each tweet into three parts (left context, target and right context), and where the sentiment towards a target entity results from the interaction between its left and right contexts. Such sentiment signal is drawn by mapping all the words in each context into low-dimensional vectors (i.e. word embeddings), using pre-trained embedding resources, and applying neural pooling functions to extract useful features. Such context set-up does not fully capture the syntactic information of the tweet and the given target entity, and by adding features from the full tweet (as done by Vo and Zhang (2015)) interactions between the left and right context are only implicitly modeled. Here we use a syntactic dependency parser designed explicitly for tweets (Kong et al., 2014) to find the syntactically connected parts of the tweet to each target. We then extract word embedding features from these syntactically dependent tokens $[D_1, ..., D_n]$ along its dependency path in the parsing tree to the target[7], as well as from the left-target-right contexts (i.e. $L - T - R$). Feature vectors generated from different contexts are concatenated into a final feature vector as shown in (2), where $P(X)$ presents a list of $k$ different pooling functions on an embedding matrix $X$. Not only does this proposed framework make the learning process efficient without labor intensive manual feature engineering and heavy architecture engineering for neural models, it has also shown that complex syntactic and semantic information can be effectively drawn by simply concatenating different types of context together without the use of deep learning (other than pre-trained word embeddings).

$$F = [P(D), P(L), P(T), P(R)]; \quad \text{with } P(X) = [f_1(X), ..., f_k(X)] \tag{2}$$

**Data set:** We evaluate and compare our proposed system to the state-of-the-art baselines on a benchmarking corpus (Dong et al., 2014) that has been used by several previous studies (Vo and Zhang, 2015; Tang et al., 2016a; Zhang et al., 2016). This corpus contains 6248 training tweets and 692 testing tweets with a sentiment class balance of 25% negative, 50% neutral and 25% positive. Although the original corpus has only annotated one target per tweet, without specifying the location of the target, we expand this notion to consider cases where the target entity may appear more than once at different locations in the tweet, e.g.:

*"**Nicki Minaj** has brought back the female rapper. - really? **Nicki Minaj** is the biggest parody in popular music since the Lonely Island."*

Semantically it is more appropriate and meaningful to consider both target appearances when determining the sentiment polarity of "Nicki Minaj" expressed in this tweet. While it isn't clear if Dong et al. (2014) and Tang et al. (2016a) have considered this realistic **same-target-multi-appearance scenario**, Vo et al. (2015) and Zhang et al. (2016) do not take it into account when extracting target-dependent contexts. Contrary to these studies we extend our system to fully incorporate the situation where a target appears multiple times at different locations in the tweet. We add another pooling layer in (2) where we apply a *medium* pooling function to combine extracted feature vectors from each target appearance together into the final feature vector for the sentiment classification of such targets. Now the feature extraction function $P(X)$ in (2) becomes:

[7]Empirically the proximity/location of such syntactic relations have not made much difference when used in feature weighting and is thus ignored.

$$P(X) = [P_{medium}([f_1(X_1), ..., f_1(X_m)]), \\ ... \ ... \ , \\ P_{medium}([f_k(X_1), ..., f_k(X_m)])] \quad (3)$$

where $m$ is the number of appearances of the target and $P_{medium}$ represents the dimension-wise *medium* pooling function.

**Models:** To investigate different ways of modelling target-specific context and evaluate the benefit of incorporating the same-target-multi-appearance scenario, we build these models:

- **Semeval-best**: is a tweet-level model using various types of features, namely ngrams, lexica and word embeddings with extensive data pre-processing and feature engineering. We use this model as a target-independent baseline as it approximates and beats the best performing system (Boag et al., 2015) in Semeval 2015 task 10. It also outperforms the highest ranking system, Tweester, on the Semeval 2016 corpus (by +4.0% in macro-averaged recall) and therefore constitutes a state-of-the art tweet level baseline.
- **Naive-seg models**: **Naive-seg-** slices each tweet into a sequence of sub-sentences by using punctuation (i.e. ',' '.' '?' '!'). Embedding features are extracted from each subsentence and pooling functions are applied to combine word vectors. **Naive-seg** extends it by adding features extracted from the left-target-right contexts, while **Naive-seg+** extends Naive-seg by adding lexicon filtered sentiment features.
- **TDParse models**: as described in Section 4.1. **TDParse-** uses a dependency parser to extract a syntactic parse tree to the target and map all child nodes to low-dimensional vectors. Final feature vectors for each target are generated using neural pooling functions. While **TDParse** extends it by adding features extracted from the left-target-right contexts, **TDParse+** uses three sentiment lexica for filtering words. **TDParse+ (m)** differs from TDParse+ by taking into account the 'same-target-multi-appearance' scenario. Both **TDParse+** and **TDParse+ (m)** outperform state-of-the-art target-specific models.
- **TDPWindow-N**: the same as **TDParse+** with a window to constrain the left-right context. For example if $N = 3$ then we only consider 3 tokens on each side of the target when extracting features from the left-right context.

### 4.2 Experimental Settings

To compare our proposed models with Vo & Zhang (2015), we have used the same pre-trained embedding resources and pooling functions (i.e. *max*, *min*, *mean*, *standard deviation* and *product*). For classification we have used LIBLINEAR (Fan et al., 2008), which approximates a linear SVM. In tuning the cost factor $C$ we perform five-fold cross validation on the training data over the same set of parameter values for both Vo and Zhang (2015)'s implementation and our system. This makes sure our proposed models are comparable with those of Vo and Zhang (2015).

**Evaluation metrics:** We follow previous work on target-dependent Twitter sentiment classification, and report our performance in accuracy, 3-class macro-averaged (i.e. negative, neutral and positive) $F_1$ score as well as 2-class macro-averaged (i.e. negative and positive) $F_1$ score [8], as used by the Semeval competitions (Rosenthal et al., 2015) for measuring Twitter sentiment classification performance.

### 4.3 Experimental results and comparison with other baselines

We report our experimental results in **Table 2** on the single-target benchmarking corpus (Dong et al., 2014), with three model categories: 1) tweet-level target-independent models, 2) target-dependent models without considering the 'same-target-multi-appearance' scenario and 3) target-dependent models incorporating the 'same-target-multi-appearance' scenario. We include the models presented in the previous section as well as models for target specific sentiment from the literature where possible.

Among the target-independent baseline models **Target-ind** (Vo and Zhang, 2015) and **Semeval-best** have shown strong performance compared with **SSWE** (Tang et al., 2014) and **SVM-ind** (Jiang et al., 2011) as they use more features, especially rich automatic features using the embeddings of Mikolov et al. (2013). Interestingly they also perform better than some of the target-dependent baseline systems, namely **SVM-dep**

[8] Note that this isn't a binary classification task; the $F_1$ score is still effected by the neutral tweets.

(Jiang et al., 2011), **Recursive NN** and **AdaRNN** (Dong et al., 2014), showing the difficulty of fully extracting and incorporating target information in tweets. Basic **LSTM** models (Tang et al., 2016a) completely ignore such target information and as a result do not perform as well.

Among the target-dependent systems neural network baselines have shown varying results. The adaptive recursive neural network, namely **AdaRNN** (Dong et al., 2014), adaptively selects composition functions based on the input data and thus performs better than a standard recursive neural network model (**Recursive NN** (Dong et al., 2014)). **TD-LSTM** and **TC-LSTM** from Tang et al. (2016a) model left-target-right contexts using two LSTM neural networks and by doing so incorporate target-dependent information. **TD-LSTM** uses two LSTM neural networks for modeling the left and right contexts respectively. **TC-LSTM** differs from (and outperforms) **TD-LSTM** in that it concatenates target word vectors with embedding vectors of each context word. We also test the Gated recurrent neural network models proposed by Zhang et al. (2016) on the same dataset. The gated models include: **GRNN**, that includes gates in its recurrent hidden layers, **G3** that connects left-right context using a gated NN structure, and a combination of the two - **GRNN+G3**. Results show these gated neural network models do not achieve state-of-the-art performance. When we compare our target-dependent model **TDParse+**, which incorporates target-dependent features from syntactic parses, against the target-dependent models proposed by Vo and Zhang (2015), namely **Target-dep** which combines full tweet (pooled) word embedding features with features extracted from left-target-right contexts and **Target-dep+** that adds target-dependent sentiment features on top of **Target-dep**, we see that our method beats both of these, without using full tweet features[9]. **TDParse+** also outperforms the state-of-the-art **TC-LSTM**.

When considering the 'same-target-multi-appearance' scenario, our best model - **TDParse+** improves its performance further (shown as **TDParse+ (m)** in Table 2). Even though **TDParse** doesn't use lexica, it shows competitive results to **Target-dep+** which uses lexicon filtered sentiment features. In the case of **TDParse-**, which uses exclusively features from syntactic parses, while it performs significantly worse than **Target-ind**, that uses only full tweet features, when the former is used in conjunction with features from left-target-right contexts it achieves better results than the equivalent **Target-dep** and **Target-dep+**. This indicates that syntactic target information derived from parses complements well with the left-target-right context representation. Clausal segmentation of tweets or sentences can provide a simple approximation to parse-tree based models (Li et al., 2015). In Table 2 we can see our naive tweet segmentation models **Naive-seg** and **Naive-seg+** also achieve competitive performance suggesting to some extent that such simple parse-tree approximation preserves the semantic structure of text and that useful target-specific information can be drawn from each segment or clause rather than the entire tweet.

| Model | Accuracy | 3 Class $F_1$ | 2 Class $F_1$ |
|---|---|---|---|
| SSWE | 62.4 | 60.5 | |
| SVM-ind | 62.7 | 60.2 | |
| LSTM | 66.5 | 64.7 | |
| Target-ind | 67.05 | 63.4 | 58.5 |
| Semeval-best | 67.6 | 64.3 | 59.2 |
| SVM-dep | 63.4 | 63.3 | |
| Recursive NN | 63.0 | 62.8 | |
| AdaRNN | 66.3 | 65.9 | |
| Target-dep | 70.1 | 67.4 | 63.2 |
| Target-dep+ | 70.5 | 68.1 | 64.1 |
| TD-LSTM | 70.8 | 69.0 | |
| TC-LSTM | 71.5 | 69.5 | |
| GRNN | 68.5 | 65.8 | 61.0 |
| G3 | 68.5 | 67.0 | 63.9 |
| GRNN+G3 | 67.9 | 65.2 | 60.5 |
| TDParse+ | **72.1** | **69.8** | **66.0** |
| Target-dep+ (m) | 70.7 | 67.8 | 63.4 |
| Naive-seg- | 63.0 | 57.6 | 51.5 |
| Naive-seg | 70.8 | 68.4 | 64.5 |
| Naive-seg+ | 70.7 | 67.7 | 63.2 |
| TDParse- | 61.7 | 57.0 | 51.1 |
| TDParse | 71.0 | 68.4 | 64.3 |
| TDParse+ (m) | **72.5** | **70.3** | **66.6** |
| TDPWindow-2 | 68.2 | 64.7 | 59.2 |
| TDPWindow-7 | 71.2 | 68.5 | 64.2 |
| TDPWindow-12 | 70.5 | 67.9 | 63.8 |

Table 2: Performance comparison on the benchmarking data (Dong et al., 2014)

## 5 Evaluating Baselines for target-specific sentiment in a multi-target setting

We perform multi-target-specific sentiment classification on our election dataset by extending

[9]Note that the results reported in Vo and Zhang (2015) (**71.1** in accuracy and **69.9** in $F_1$) were not possible to reproduce by running their code with very fine parameter tuning, as suggested by the authors

and applying our models described in Section 4.1. We compare the results with our other developed baseline models in Section 4.1, including a tweet-level model **Semeval-best** and clausal-segmentation models that provide simple parse-tree approximation, as well as state-of-the-art target-dependent models by Vo and Zhang (2015) and Zhang et al. (2016). The experimentation set-up is the same as described in Section 4.2[10].

**Data set:** Our election data has a training/testing ratio of 3.70, containing 3210 training tweets with 9912 target entities and 867 testing tweets with 2675 target entities.

**Models:** In order to limit our use of external resources we do not include **Naive-seg+** and **TDParse+** for evaluation as they both use lexica for feature generation. Since most of our tweets here contain $N > 1$ targets and the target-independent classifiers produce a single output per tweet, we evaluate its result $N$ times against the ground truth labels, to make different models comparable.

**Results:** Overall the models perform much poorer than for the single-target benchmarking corpus, especially in 2-class $F_1$ score, indicating the challenge of the multi-target-specific sentiment recognition. As seen in Table 3 though the feature-rich tweet-level model **Semeval-best** gives a reasonably strong baseline performance (same as in Table 2), both it and **Target-ind** perform worse than the target-dependent baseline models **Target-dep/Target-dep+** (Vo and Zhang, 2015), indicating the need to capture and utilise target-dependent signals in the sentiment classification model. The Gated neural network models - **G3/GRNN/GRNN+G3** (Zhang et al., 2016) also perform worse than **Target-dep+** while the combined model - **GRNN+G3** fails to boost performance, presumably due to the small corpus size.

Our final model **TDParse** achieves the best performance especially in 3-class $F_1$ and 2-class $F_1$ scores in comparison with other target-dependent and target-independent models. This indicates that our proposed models can provide better and more balanced performance between precision and recall. It also shows the target-dependent syntactic information acquired from parse-trees is beneficial to determine the target's sentiment particularly when used in conjunction with the left-target-right contexts originally proposed by Vo and Zhang (2015) and in a scenario of multiple targets per tweet. Our clausal-segmentation baseline - **Naive-seg** models approximate such parse-trees by identifying segments of the tweet relevant to the target, and as a result **Naive-seg** achieves competitive performance compared to other baselines.

[10]Class weight parameter is not optimised for all experiments, though better performances can be achieved here by tuning the class weight due to the class imbalance nature of this dataset.

| Model | Accuracy | 3 Class $F_1$ | 2 Class $F_1$ |
|---|---|---|---|
| Semeval-best | 54.09 | 42.60 | 40.73 |
| Target-ind | 52.30 | 42.19 | 40.50 |
| Target-dep | 54.36 | 41.50 | 38.91 |
| Target-dep+ | 55.85 | 43.40 | 40.85 |
| GRNN | 54.92 | 41.22 | 38.57 |
| G3 | 55.70 | 41.40 | 37.87 |
| GRNN+G3 | 54.58 | 41.04 | 39.46 |
| Naive-seg- | 51.89 | 39.94 | 37.17 |
| Naive-seg | 55.07 | 43.89 | 40.69 |
| TDParse- | 52.53 | 42.71 | 40.67 |
| TDParse | 56.45 | **46.09** | **43.43** |
| TDPWindow-2 | 55.10 | 43.81 | 41.36 |
| TDPWindow-7 | 55.70 | 44.66 | 41.35 |
| TDPWindow-12 | **56.82** | 45.45 | 42.69 |

Table 3: Performance comparison on the election dataset[11]

| S1 | Semeval-best | Target-dep+ | TDParse |
|---|---|---|---|
| Macro 3-class-F1 | 50.11 | 46.24 | 47.08 |
| Micro 3-class-F1 | 59.72 | 55.82 | 57.47 |
| Macro 2-class-F1 | 46.59 | 43.42 | 42.95 |
| **S2** | **Semeval-best** | **Target-dep+** | **TDParse** |
| Macro 3-class-F1 | 37.15 | 41.81 | 43.07 |
| Micro 3-class-F1 | 45.17 | 51.66 | 52.05 |
| Macro 2-class-F1 | 37.05 | 39.75 | 40.92 |
| **S3** | **Semeval-best** | **Target-dep+** | **TDParse** |
| Macro 3-class-F1 | 35.08 | 42.83 | 51.26 |
| Micro 3-class-F1 | 38.16 | 46.05 | 53.07 |
| Macro 2-class-F1 | 35.17 | 40.53 | 50.14 |

Table 4: Performance analysis in **S1**, **S2** and **S3**

### 5.1 State-of-the-art tweet level sentiment vs target-specific sentiment in a multi-target setting

To fully compare our multi-target-specific models against other target-dependent and target-independent baseline methods, we conduct an additional experiment by dividing our election data test set into three disjoint subsets, on the basis of number of distinct target sentiment values per tweet: (**S1**) contains tweets having only one target sentiment, where the sentiment towards each target is the same; (**S2**) and (**S3**) contain two and three different types of targeted sentiment respec-

[11]Any further results will be shared on our Github page: https://goo.gl/S2T1GO

tively (i.e. in **S3**, positive, neutral and negative sentiment are all expressed in each tweet). As described in Section 3.2, there are 2,051, 1,753 and 273 tweets in S1, S2 and S3 respectively.

Table 4 shows results achieved by the tweet-level target-independent model - **Semeval-best**, the state-of-the-art target-dependent baseline model - **Target-dep+**, and our proposed final model - **TDParse**, in each of the three subsets. We observe **Semeval-best** performs the best in **S1** compared to the two other models but its performance gets worse when different types of target sentiment are mentioned in the tweet. It has the worst performance in **S2** and **S3**, which again emphasises the need for multi-target-specific sentiment classification. Finally, our proposed final model **TDParse** achieves better performance than **Target-dep+** consistently over all subsets indicating its effectiveness even in the most difficult scenario **S3**.

## 6 Conclusion and Future work

In this work we introduce the challenging task of multi-target-specific sentiment classification for tweets. To help the study we have generated a multi-target Twitter corpus on UK elections which will be made publicly available. We develop a state-of-the-art approach which utilises the syntactic information from parse-tree in conjunction with the left-right context of the target. Our method outperforms previous approaches on a benchmarking single-target corpus as well as our new multi-target election data. Future work could investigate sentiment connections among all targets appearing in the same tweet as a multi-target learning task, as well as a hybrid approach that applies either Semeval-best or TDParse depending on the number of targets detected in the tweet.

## Acknowledgments

We would like to thank Duy-Tin Vo, Meishan Zhang and Duyu Tang for sharing their implementation code respectively, which we have used for system performance comparison. We would also like to thank Li Dong for sharing their data, and City University London for recruiting PhD students for the annotation of our election corpus.

## References

Steven Bird. 2006. Nltk: the natural language toolkit. In *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, pages 69–72, Stroudsburg, PA, USA. Association for Computational Linguistics.

William Boag, Peter Potash, and Anna Rumshisky. 2015. Twitterhawk: A feature bucket based approach to sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 640–646, Denver, Colorado, June. Association for Computational Linguistics.

Pete Burnap, Rachel Gibson, Luke Sloan, Rosalynd Southern, and Matthew Williams. 2016. 140 characters to victory?: Using twitter to predict the uk 2015 general election. *Electoral Studies*, 41:230–233.

Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 49–54, Baltimore, Maryland, June. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Jenny Rose. Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 363–370, Ann Arbor, Michigan, June. Association for Computational Linguistics.

Long Jiang, Mo Yu, Ming Zhou, Xiaohua Liu, and Tiejun Zhao. 2011. Target-dependent twitter sentiment classification. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 151–160, Portland, Oregon, USA, June. Association for Computational Linguistics.

Lingpeng Kong, Nathan Schneider, Swabha Swayamdipta, Archna Bhatia, Chris Dyer, and Noah A. Smith. 2014. A dependency parser for tweets. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1001–1012, Doha, Qatar, October. Association for Computational Linguistics.

J Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174.

Jiwei Li, Thang Luong, Dan Jurafsky, and Eduard Hovy. 2015. When are tree structures necessary

for deep learning of representations? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2304–2314, Lisbon, Portugal, September. Association for Computational Linguistics.

Rada Mihalcea. 2004. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *The Companion Volume to the Proceedings of 42st Annual Meeting of the Association for Computational Linguistics*, pages 170–173, Barcelona, Spain, July. Association for Computational Linguistics.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. Semeval-2016 task 4: Sentiment analysis in twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1–18, San Diego, California, June. Association for Computational Linguistics.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Shubham Pateria and Prafulla Choubey. 2016. AKTSKI at semeval-2016 task 5: Aspect based sentiment analysis for consumer reviews. In *Proceedings of the 10th International Workshop on Semantic Evaluation, SemEval@NAACL-HLT 2016, San Diego, CA, USA, June 16-17, 2016*, pages 318–324.

Nataliia Plotnikova, Micha Kohl, Kevin Volkert, Stefan Evert, Andreas Lerner, Natalie Dykes, and Heiko Ermer. 2015. Klueless: Polarity classification and association. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 619–625, Denver, Colorado, June. Association for Computational Linguistics.

Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. 2015. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June. Association for Computational Linguistics.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1524–1534, Stroudsburg, PA, USA. Association for Computational Linguistics.

Sara Rosenthal, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov. 2015. Semeval-2015 task 10: Sentiment analysis in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 451–463, Denver, Colorado, June. Association for Computational Linguistics.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland, June. Association for Computational Linguistics.

Duyu Tang, Bing Qin, Xiaocheng Feng, and Ting Liu. 2016a. Effective lstms for target-dependent sentiment classification. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3298–3307, Osaka, Japan, December. The COLING 2016 Organizing Committee.

Duyu Tang, Bing Qin, and Ting Liu. 2016b. Aspect level sentiment classification with deep memory network. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 214–224, Austin, Texas, November. Association for Computational Linguistics.

Richard Townsend, Adam Tsakalidis, Yiwei Zhou, Bo Wang, Maria Liakata, Arkaitz Zubiaga, Alexandra Cristea, and Rob Procter. 2015. Warwickdcs: From phrase-based to target-specific sentiment recognition. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 657–663, Denver, Colorado, June. Association for Computational Linguistics.

Andranik Tumasjan, Timm Oliver. Sprenger, Philipp G. Sandner, and Isabell M. Welpe. May 2010. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185.

Saúl Vargas, Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2016. Comparing overall and targeted sentiments in social media during crises. In *Proceedings of the Tenth International Conference on Web and Social Media, Cologne, Germany, May 17-20, 2016.*, pages 695–698.

Duy-Tin Vo and Yue Zhang. 2015. Target-dependent twitter sentiment classification with rich automatic features. In *Proceedings of the 24th International Conference on Artificial Intelligence*, IJCAI'15, pages 1347–1353. AAAI Press.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *Proceedings of the ACL 2012 System Demonstrations*, pages 115–120,

Jeju Island, Korea, July. Association for Computational Linguistics.

Meishan Zhang, Yue Zhang, and Duy-Tin Vo. 2016. Gated neural networks for targeted sentiment analysis. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, pages 3087–3093. AAAI Press.