Fast Approach to Build an Automatic Sentiment Annotator for Legal Domain using Transfer Learning

Viraj Salaka Gamage, Menuka Warushavithana, Nisansa de Silva,
Amal Shehan Perera, Gathika Ratnayaka and **Thejan Rupasinghe**
Department of Computer Science & Engineering
University of Moratuwa
viraj.14@cse.mrt.ac.lk

Abstract

This study proposes a novel way of identifying the sentiment of the phrases used in the legal domain. The added complexity of the language used in law, and the inability of the existing systems to accurately predict the sentiments of words in law are the main motivations behind this study. This is a transfer learning approach which can be used for other domain adaptation tasks as well. The proposed methodology achieves an improvement of over 6% compared to the source model's accuracy in the legal domain.

1 Introduction

As described by Esuli and Sebastiani (2007), sentiment analysis or *sentiment classification* is a recent methodology that aligns with information retrieval and computational linguistics which is focused on the opinion towards something which is represented by a certain text.

In many recent studies involving NLP in various domains, it is common to reuse the seminal RNTN (Recursive Neural Tensor Network) model (Socher et al., 2013b) trained on movie reviews for sentiment analysis. However, the trained model has bias towards the movie review domain.

We propose a novel methodology to perform transfer learning on the RNTN model mentioned in Socher et al. (2013b) and build a target model. Given that this is a transfer learning approach, the manually annotated data on movie reviews is used as the initial source model, rather than creating a new comparable manually annotated dataset for the legal domain.

In the proposed approach, the sentiment of a given phrase is classified into one of the two classes; *negative* and *non-negative*. This classification criterion is selected following the fact that the major use case aligns with classifying terms and entities supporting/referring to either plaintiff or defendant. Therefore, the proposed methodology is focused on identifying the statements with negative sentiment as much as possible. This kind of sentiment classification is vital to identify the stakeholder-bias in legal case statements. Similarly, sentiment analysis in legal text can become useful in automating the identification of arguments, the supporting/opposing party for a given argument and counter arguments.

For the testing purposes, we created a manually annotated target domain test dataset such that the phrases belong to one of the two classes: *negative* or *non-negative*. The target system shows a recall of 0.7014 for identifying phrases with negative sentiment in the legal domain. Furthermore, the overall accuracy of the system is above 76% in classifying sentiments for a given phrase correctly. If this result is compared with the results of source RNTN model (Socher et al., 2013b), it is a 6% improvement in accuracy. The approach proposed in this study can be tried on other domain adaptation tasks related to sentiment classification as well.

2 Background

The legal vocabulary have words of mixed origin such as English and Latin has been raised as a reason for the difficulty of creating computing applications for the legal domain (Sugathadasa et al., 2018). However, recently, there have been attempts to involve and build legal ontologies (Jayawardana et al., 2017a,b,c). Given the popularity of knowledge embedding, a number of studies have also attempted to embed legal jargon in vector spaces (Sugathadasa et al., 2017; Nay, 2016). If we consider the research on sentiment analysis in legal domain, the study on *Opinion Mining* in legal blogs (Conrad and Schilder, 2007) is closest implementation for this study that we

Proceedings of the 9th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pages 260–265

https://doi.org/10.18653/v1/P17

have found. But, the data set used for evaluation is based on movie reviews, customer reviews, and MPQA corpus (Wiebe et al., 2005).

There have been numerous studies that were built upon SentiWordNet (Esuli and Sebastiani, 2007; Baccianella et al., 2010) which attempts to classify sentiments of phrases and sentences. One such study by Ohana and Tierney (2009) proposes a methodology to perform opinion mining on movie reviews using support vector machine where some of the features were calculated using WordNet. This achieves an accuracy of 69.35% and claims that the inaccuracies in SentiWordNet feature calculations are caused by the SentiWordNet's reliance on glosses. Lu et al. (2012) evaluates the SentiWordNet for identifying opposing opinion networks in forum discussion. The average SentiWordNet opinion score of words is considered to identify whether a user's expressed comment for a given post has either *for* or *against* relationship. The achieved accuracy using the SentiWordNet opinion score of words is 0.56.

The method proposed by Socher et al. (2013b) provides an algorithm to identify the sentiment of a phrase or a sentence in a supervised manner using a deep learning model of the type Recursive Neural Tensor Network (RNN). It is claimed that this learning model has the capability to identify the sentiment considering the context of that word. A dataset which consists of movie reviews where each sentence in the data set was broken into phrases and each phrase is annotated by human judges were created for this study. The authors claim a testing accuracy of 80.7% in phrase level for a test set drawn from the same dataset. Further, the authors claim that the proposed model can be trained over any domain by following the provided methodology. While, theoretically, it is possible, following this for legal domain in a practical implementation which covers a corpus which is both significant and sufficient is difficult. This claim is substantiated by referring the dataset of the original research (Socher et al., 2013b) which utilized 215,154 manually annotated phrases (from 11,855 sentences) with over 5355 unique words. In comparison to this, the legal corpus used in our study has a vocabulary exceeding 17000 words. The difficulties are not mealy of scale given that the linguistic complexity of legal jargon exceeds that of the average text corpus (Jayawardana et al., 2017b,c; Sugathadasa et al., 2017, 2018).

Domain adaptation is a sub-category of *Transfer Learning* (Raina et al., 2007). There are several studies (Raina et al., 2007; Socher et al., 2013a) that claim the process of *domain adaptation* to be a suitable solution to perform transfer learning. While the generic process of transfer learning is defined as the process of "learning model is trained using data from a certain domain and tested with respect to a different domain" (Raina et al., 2007), the specific case of *domain adaptation* occurs when the task is similar in both source and target models. Quattoni et al. (2008) is a study based on domain adaptation in Image Classification.

3 Methodology

Given that the transfer learning process described in this study uses the Recursive Neural Tensor Network (RNTN) model proposed by Socher et al. (2013b) as the source model, we make numerous references to the aforementioned model throughout the paper. Therefore, to avoid clutter, from this point onward the model proposed by Socher et al. (2013b) is referred as **Socher Model** in the remainder of this paper.

3.1 Selecting the Vocabulary

Depending on the size of the corpus (phrases extracted from legal text), availability of human annotators and the time, it is not feasible to analyze and modify the sentiment of every word in a corpus. Therefore, it is required to select the vocabulary (unique words in the corpus) such that the end-model can correctly classify the sentiment of most of the phrases from the legal domain while not squandering human annotator time on words that occur rarely. To this end, first, the stop-words (Lo et al., 2005) are removed from the text by utilizing the classical stop-word list known as the Van stop-list (Van Rijsbergen, 1979). Next, the term frequencies for each word in the corpus is calculated and only the top 95% words of it are added to the vocabulary.

3.2 Assigning Sentiments for the Selected Vocabulary

The selected vocabulary (set of individual words) is given to the sentiment annotator *Socher Model* as input. From the model, sentiment is classified into one of the five classes as in table 3.2. This class scheme made sense for the movie re-

views for which the *Socher Model* is trained and used for. However, in the application of this study, the basic requirement of finding sentiment in *court cases* in the legal domain is to identify whether a given statement is against the plaintiff's claim or not. Therefore, we define two classes for sentiment: *negative* and *non-negative*.

Three human judges analyze the selected vocabulary and classify each unique word into the two classes depending on its sentiment separately and independently. If at least two judges agree, the given word's sentiment is assigned as the class those two judges agreed. For the same word, the output from the sentiment annotator *Socher Model* belongs to one of the five classes mentioned in the preceding subsection. In this approach, we map the output from *Socher Model* to the two classes we define in Table 3.2.

	Human annotation	***Socher Model* output**
Class 1	Negative	Very negative, negative
Class 2	Non-negative	Neutral, Positive, very positive

Table 1: Sentiment Mapping

For a given word, if the two sentiment values assigned by the *Socher Model* and human judges do not agree with the above mapping, we define that the *Socher Model*'s output has deviated from its actual sentiment. For example:

Sentence: *Sam is charged with a crime.*
***Socher Model*'s output:** positive
Human judges' annotation: negative

The word *charged* has several meanings depending on the context. As the *Socher Model* was trained using movie reviews, the sentiment of the word *charged* is identified as positive. Although the sentiment of the term *crime* is recognized as negative, the sentiment of the whole sentence is output as positive. But in the legal domain, *charged* refers to a formal accusation. Therefore, the sentiment for the above sentence should have been negative. From the selected vocabulary, all the words with deviated sentiments are identified and listed separately for the further processing.

3.3 Brief description on the RNTN Model

In the preceding subsection, we came across a situation where the sentiment values from the *Socher Model* do not match the actual sentiment value because of the difference in domains. And there are words like *insufficient*, which were not recognized by the model because those terms were not included in the training data-set. One approach to solve this is to annotate the phrases extracted from legal case transcripts manually as the *Socher Model* suggests, which will require a considerable amount of human effort and time. Instead of that, we can change the model such that the desired output can be obtained using the same trained *Socher Model* without explicitly training using phrases in the legal domain. Hence, this method is called a transfer learning method.

In order to change the model, first, it is required to understand the internals of the *Socher Model* model. When a phrase is provided as input, first it generates a binary tree corresponding to the input in which each leaf node represents a single word. Each leaf node is represented as a vector with d-dimensions. The parent nodes are also d-dimensional vectors which are computed in the bottom-up fashion according to some function g. The function g is composed of a neural tensor layer. Through the training process, the neural tensor layer and the word vectors are adjusted to support the relevant sentiment value. The neural tensor layer corresponds to identify the sentiment according to the structure of words representing the phrase. If we consider a phrase like *not guilty* ,both individual word elements have negative sentiments. But the composition of those words has the structure of negating a negative sentiment term or phrase. Hence the phrase has a non-negative sentiment. If the input was a phrase like *very bad*, the neural tensor layer has the ability to identify that the term *very* increases the negativity in the sentiment.

3.4 Adjusting Word Vector Values in RNTN Model

The requirement of the system is to identify the sentiment of a given phrase. The proposed approach is not to modify the neural tensor layer completely. We simply substitute the word vector values of individual words which are having deviated sentiments between *Socher Model* and human annotation (See sections 3.2). The vectors for the words which were not in the vocabulary of the training set which was used to train the RNTN model should be instantiated. The vectors of the words which are not deviated (according to the definition provided in the preceding subsection

3.3) will remain the same.

As the words with deviated sentiments (provided by the *Socher Model*) in the vocabulary are already known, we initialize the vectors corresponding to the sentiment annotation for those words. Since the model is not trained explicitly, the vector initialization is done by substituting the vectors of words in which sentiment is not deviated comparing the *Socher Model* output and its actual sentiment. After the substitution is completed, we consider the part-of-speech tag. For that purpose, the part-of-speech tagger mentioned in Toutanova et al. (2003) is used. The substitution of vectors is carried out as shown in Table 2.

POS Tag	Substituted word vector sentiment	
	non-negative	negative
NN	failure	thing
RB	insufficiently	naturally
VB	hate	do
VBZ	ignoring	doing

Table 2: Substituted Word Vectors for words which should be deviated

The number of words which have deviated sentiments is a considerably lower amount compared to the selected vocabulary. The rest of the words' vectors representing sentiments are not changed in the modification process. The neural tensor layer also remains unchanged from the trained *Socher Model* using movie reviews (Socher et al., 2013b). When the vectors for words with deviated sentiments are initialized according to the part-of-speech tag as shown in Table 2, it is possible to make a fair assumption that when deciding the sentiment with the proposed implementation, it does not harm the structure corresponding to the linguistic features of English. Consider the sentence "*evidence is insufficient.*" as an example.

The term "*insufficient*" is not in the vocabulary of the *Socher Model* due to the limited vocabulary in training data set. Therefore, the *Socher Model* provides the sentiment of that word as neutral which indicates as a word with a deviated sentiment. Following the Table 2, the sentiment related vector is instantiated by substituting the vector of **wrong** as the part-of-speech tag of **insufficient** is **JJ** (Santorini, 1990). Therefore the modified version of the RNTN model has the capability of identifying the sentiment of the above sentence as negative. The figure 1 shows how the sentiment is induced through the newly instantiated word vector.

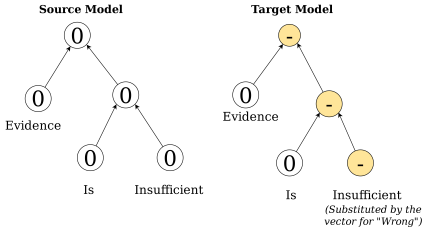

Figure 1: Sentiment Prediction for a phrase with words not in source's vocabulary but in target's vocabulary

And there are scenarios where the term is in the vocabulary of the *Socher Model* but has a different sentiment compared to the legal domain. Consider the sentence "*Sam is charged with a crime*" which was mentioned in section 3.2,

In section 3.2, we have identified that the term *charged* denotes a different sentiment in legal domain compared to movie reviews. The source RNTN model outputs a positive sentiment for that given sentence as the term *charged* is identified as having a positive sentiment according to movie reviews domain. And that term is the cause for having such an output from the source model. The figure 2 indicates how the change we introduced in the target model (in section 3.2) induce the correct sentiment up to the root level of the phrase. Therefore, the target model identifies the sentiment correctly for the given phrase.

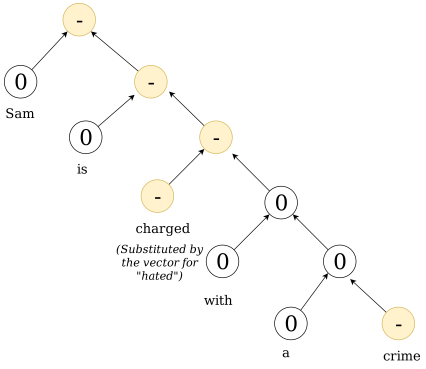

Figure 2: Sentiment Prediction for a phrase with words having deviated sentiment in two domains - target model

To improve the recall in identifying phrases with negative sentiment, we have added another rule to the classification criteria. The source RNTN model (*Socher Model*) provides the score for each of the five classes such that all those five scores sum up to 1. If the negative sentiment class has the highest score, the sentiment label of the

phrase will be *negative*. Otherwise, the phrase again can be classified as having a *negative* sentiment if the score for negative sentiment class is above 0.4. If those two conditions are not met, the phrase will be classified as having a *non-negative* sentiment. Section 4 provides observations and results regarding the improved criteria.

4 Experiments and Results

The proposed approach in this paper is based on transfer learning. Therefore, we needed to create a golden standard for identifying sentiments of phrases and sentences in the legal domain in order to evaluate the model. The phrases and sentences for the test data set are randomly picked from legal case transcripts based on the United States Supreme Court. During the selection process, we have selected an equal amount of phrases for both classes according to the *Socher Model*. Each of these phrases and sentences is annotated by three human annotators. Since the classification process is binary, we pick the sentiment class for each test subject based on the maximum number of votes. In the end, we prepare the test data set containing nearly 1500 annotations to use in the evaluation process.

In the experiment, we compare the sentiment class picked by human judges and the modified RNTN model. As the baseline model, we use the source RNTN model (*Socher Model*) to check the impact caused by the proposed transfer learning approach. The acquired results from the baseline model is shown in Table 3 and results from the target model is shown in Table 4.

According to Table 3 and Table 4, there is a 10% improvement in identifying phrases with negative sentiment. The reason is that there are a lot of unknown words which are in the legal domain but not in movie reviews corpus. In addition, we have introduced new criteria based on a threshold for the score of negative class to improve the recall. Due to that reason, the precision in identifying phrases with a negative sentiment is 0.8441. But if we compare with the precision of the baseline model (*Socher Model*) for negative sentiment class is 0.7962 which is a lower value. Since the test dataset is not skewed a lot towards one class, it is fair to consider the accuracy of the system in predicting the sentiment for any given phrase. The baseline model shows the accuracy of 70.17% while the target model shows 76.80%. The improvement in accuracy is above 6%.

Actual \ Predicted	**Negative**	**Non-negative**	***Total***
Negative	60.43%	39.57%	278
Non-negative	18.29%	81.71%	235
Total	211	301	513

Table 3: Confusion Matrix for Results from the Baseline Model

Actual \ Predicted	**Negative**	**Non-negative**	***Total***
Negative	70.14%	29.86%	278
Non-negative	15.32%	84.68%	235
Total	231	282	513

Table 4: Confusion Matrix for Results from the Improved Model

The observed results in Table 3 and Table 4 show that there is a 6% improvement of the sentiment with respect to the baseline model. There are a few reasons behind the results. As we randomly selected phrases from the legal case transcripts corpus, only 45% of the phrases actually contained the words where we had substituted the vector regarding sentiment. Therefore, the output for 55% of the phrases from the baseline model and the target model was the same. If we compare the output provided by the baseline model and the target model, output of 9.5% of the total phrases are different to each other. Therefore the difference between the two models is based on that 9.5% of the total phrases.

5 Conclusion

This study is focused on building an automatic sentiment annotator for legal texts based on the *Recursive Neural Tensor Network (RNTN)* model mentioned in Socher et al. (2013b). Furthermore, this study can be identified as a transfer learning approach as it is not required to prepare a training data set for the legal domain specifically. Instead, this approach uses the same training data set stated in Socher et al. (2013b). This task can be recognized as a domain adaptation task. The proposed approach could achieve a 70.14% recall in identifying phrases with negative sentiments (improvement is 10% compared to the source model). The accuracy of the target model is above 76% which is a 6% improvement over the source model.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *Lrec*, volume 10, pages 2200–2204.

Jack G Conrad and Frank Schilder. 2007. Opinion mining in legal blogs. In *Proceedings of the 11th international conference on Artificial intelligence and law*, pages 231–236. ACM.

Andrea Esuli and Fabrizio Sebastiani. 2007. Sentiwordnet: a high-coverage lexical resource for opinion mining. *Evaluation*, 17:1–26.

Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Keet Sugathadasa, and Buddhi Ayesha. 2017a. Deriving a representative vector for ontology classes with instance word vector embeddings. In *Innovative Computing Technology (INTECH), 2017 Seventh International Conference on*, pages 79–84. IEEE.

Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Keet Sugathadasa, Buddhi Ayesha, and Madhavi Perera. 2017b. Semi-supervised instance population of an ontology using word vector embedding. In *Advances in ICT for Emerging Regions (ICTer), 2017 Seventeenth International Conference on*, pages 1–7. IEEE.

Vindula Jayawardana, Dimuthu Lakmal, Nisansa de Silva, Amal Shehan Perera, Keet Sugathadasa, Buddhi Ayesha, and Madhavi Perera. 2017c. Word vector embeddings and domain specific semantic based semi-supervised ontology instance population. *International Journal on Advances in ICT for Emerging Regions*, 10(1):1.

Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, volume 5, pages 17–24.

Yue Lu, Hongning Wang, ChengXiang Zhai, and Dan Roth. 2012. Unsupervised discovery of opposing opinion networks from forum discussions. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 1642–1646. ACM.

John J Nay. 2016. Gov2vec: Learning distributed representations of institutions and their legal text. *arXiv preprint arXiv:1609.06616*.

Bruno Ohana and Brendan Tierney. 2009. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13.

Ariadna Quattoni, Michael Collins, and Trevor Darrell. 2008. Transfer learning for image classification with sparse prototype representations. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. 2007. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766. ACM.

Beatrice Santorini. 1990. Part-of-speech tagging guidelines for the penn treebank project (3rd revision). *Technical Reports (CIS)*, page 570.

Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. 2013a. Zero-shot learning through cross-modal transfer. In *Advances in neural information processing systems*, pages 935–943.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013b. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2017. Synergistic union of word2vec and lexicon for domain specific semantic similarity. In *Industrial and Information Systems (ICIIS), 2017 IEEE International Conference on*, pages 1–6. IEEE.

Keet Sugathadasa, Buddhi Ayesha, Nisansa de Silva, Amal Shehan Perera, Vindula Jayawardana, Dimuthu Lakmal, and Madhavi Perera. 2018. Legal document retrieval using document vector embeddings and deep learning. *arXiv preprint arXiv:1805.10685*.

Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics.

CJ Van Rijsbergen. 1979. Information retrieval. dept. of computer science, university of glasgow. *URL: citeseer. ist. psu. edu/vanrijsbergen79information. html*, 14.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210.