

Inducing Distant Supervision in Suggestion Mining through Part-of-Speech Embeddings

Sapna Negi

Insight Centre for Data Analytics
Galway
sapna.negi@insight-centre.org

Paul Buitelaar

Insight Centre for Data Analytics
Galway
paul.buitelaar@insight-centre.org

ABSTRACT

Suggestion Mining refers to the task of identifying suggestion expressing sentences in a given text, which is primarily a binary sentence classification task. Suggestion mining is a young problem as compared to the other well established text classification tasks, and thus lacks large hand labeled benchmark datasets. In this work, two approaches for distant supervision are proposed, where both of them use a large silver standard dataset which we refer to as the Wiki Suggestion dataset. This dataset is constructed from sentences extracted from the articles on wikiHow and Wikipedia, serving as positive and negative instances respectively. Both the approaches use a Long Short Term Memory network based architecture to learn the classification model, but vary in their method to use the silver standard dataset. The first approach directly trains the classifier on this dataset, while the second approach only learns word level representations from this dataset. In the second approach, in addition to learning word vectors, we also learn vectors for Part of Speech (POS) tags. Interestingly, using these POS representations as input features to the neural network classifiers yields the best classification accuracy despite of their very small vocabulary as compared to the word vectors.

KEYWORDS

Suggestion Mining, Sentence Classification, Distant Supervision, Representation Learning

1 INTRODUCTION

Sentences expressing advice, tips, and recommendations can often be found among the opinionated texts like reviews, blogs, tweets, discussions etc (see Table 1). Such sentences can be collectively referred to as suggestions [7]. With the increasing availability of opinionated text, methods for automatic extraction of suggestions can be employed in different applications. The automatic extraction of suggestions from a given text is referred to as suggestion mining. Although humans can infer suggestions from any given informative text, automatic inferring of suggestions might still be a far fetched task for machines. Therefore, suggestion mining has been approached with a focus on detecting the existing sentences which express suggestions in an explicit manner [7]. Studies performed in the past have defined suggestion mining as a sentence classification task, where class prediction has to be made on each sentence of a given opinionated text, classes being *suggestion* and *non suggestion*. State of the art opinion mining systems primarily have mostly focused on identifying sentiment polarity of the text. Suggestion mining remains a very less explored area as compared to Sentiment

Source	Sentence
Hotel reviews	Be sure to specify a room at the back of the hotel.
Electronics reviews	I would recommend doing the upgrade to be sure you have the best chance at trouble free operation.
Discussion thread	If you do book your own airfare, be sure you don't have problems if Insight has to cancel the tour or reschedule it
Tweets	Dear Microsoft, release a new zune with your wp7 launch on the 11th. It would be smart.

Table 1: Examples of suggestions from different sources of opinions

Analysis, specially in the context of recent advancements in neural network based approaches for learning feature representations, which is the primary focus of this work.

In a standard use of language, suggestions are expressed with specific grammatical properties, often accompanied with keywords like *advice*, *suggest*, *recommend* etc. However, informal and figurative use of language on the web results in a limited performance of the manually constructed rule based classifiers [6]. We observe that suggestions can be a complex combination of semantic and syntactic features. Sometimes, the features are also dependent on the domain or source of data from where suggestions are extracted, because specific topics are highly likely to appear in suggestions from a given domain. Consider two sentences from hotel reviews, *You must take a long and gloomy corridor, with no decoration, which looks like you are in a basic motel.*, and *You must visit the Ice Bar, 15 euros and includes a drink*. Former is annotated as a non-suggestion and latter as a suggestion, however they possess similar grammatical properties. In a statistical approach, given the domain specific training data, a classifier can model that suggestions are more likely to talk about a bar, than a corridor.

Neural networks are proving to be highly effective in learning automatic representations of data which is best suited for the task at hand. The existing manually labeled training datasets available for suggestion mining do not seem to be large enough to learn such representations. The problem alleviates when domain specific datasets are not available, since topics and class distribution of suggestions vary with domains. Suggestion sentences in the hand labeled datasets developed by the previous works are very sparse, ranging from 8% to 27% of the total annotated sentences across

different datasets [6]. Therefore, we looked for existing sources of suggestion expressing sentences, and discovered that wikiHow (see Figure 2) could be one potential source. In this work, we propose and evaluate two approaches of employing a large silver standard dataset assembled using the text from wikiHow and Wikipedia, which we refer to as the *Wiki Suggestion Dataset*. An open domain evaluation of the approaches is also performed, where the evaluation datasets cover domains which are different from the gold standard dataset used for training the classifier. The contributions of this work can be summarized as:

- (1) Development of a large silver standard training dataset using wikiHow and Wikipedia.
- (2) Improving the classifier performance over the baseline methods, using the proposed approaches for leveraging the silver standard dataset.
- (3) Experimentation with different variations of the proposed approaches providing interesting findings. Most significant one is the improved classification performance in the experiments where the words in the datasets are replaced by their POS tags, and 50 dimensional POS embeddings with a very small vocabulary are employed.

2 RELATED WORK

2.1 Suggestion Mining

The previous approaches for suggestion mining include linguistic rules [2, 10], and supervised machine learning with manually determined feature types. The main algorithms used for supervised learning were, HMM and CRF [14], Factorisation Machines [3], and Support Vector Machines[7]. These statistical classifiers were mostly trained and evaluated on datasets from a single domain. These studies also provided training datasets, which were all smaller than 8000 sentences, with a highly imbalanced class distribution. Only some of these datasets are publicly available. Suggestion class was the minority class in all of these datasets, with its proportion varying from 8% to 27% of all the sentences in the dataset.

Recently, Negi et al. [6] stressed upon the use of neural networks for suggestion mining, and used pre-trained word embeddings with the gold standard training datasets. They compared different classifiers, including manually formulated rules, SVM with a variety of manually defined lexical, syntactic, and sentiment features, Convolutional Neural Networks, and LSTM Networks. They used 400 dimensional pre-trained embeddings, which were trained using best performing configurations for the Word2Vec algorithm [1]. They performed cross-validation on a number of datasets from different domains, and their results show that LSTM Networks perform consistently better with majority of the datasets, as compared to the other methods. They also performed cross domain training and testing and the neural network based approaches using pre-trained word embeddings as input representations. Based on these findings, we also employ a LSTM based classification architecture in this work, and use the approach from Negi et al. [6] as baseline.

2.2 Distant Supervision in Suggestion Mining

To the best of our knowledge, only one study related to suggestion mining has focused on using distant supervision for extracting

suggestions for improvement from customer feedback [5]. This study first defines a set of what they refer to as *lexical-Part of Speech* patterns to identify suggestion reviews in a large dataset of about 50,000 eBay App reviews on App store and Google Play. The predicted suggestions were then used as positive instances to train a SVM based classifier which used bag of words features. They performed distant supervision at the review level, rather than the sentence level. They compared the results with a SVM classifier trained on a manually labeled dataset. The classifier trained on manually labeled reviews obtained higher precision, recall and F-measure values (0.38, 0.78, 0.51 respectively) on a test dataset, compared to their distant supervision method (0.32, 0.74, and 0.46 respectively). In order to extract the exact suggestion sentence within the reviews identified in the first stage, they simply used a SVM classifier with bag of words features trained on a manually annotated sentence dataset. None of their datasets are available for open research.

To the best of our knowledge, this work is the first to use distant supervision for suggestion mining by means of an open domain silver standard dataset and representation learning. In addition, we evaluate our approach for domain independent training.

2.3 Learning Continuous Representations of Words

Pennigton et al. [8] provided GloVe algorithm to train general purpose word embeddings which were shown to outperform other high performance algorithms on various benchmark tasks and datasets. The GloVe outperformed algorithms included the algorithms like *skip grams* and *CBOW* which are two variations of the popular *word2vec* model. Therefore, the pre-trained GloVe embeddings used in [8] are a strong baseline to evaluate the performance of the embeddings learned using the wiki suggestion dataset.

Training task specific embeddings have been proven to be useful for other short text classification tasks like sentiment analysis. Tang et al. [11] trained sentiment specific word embeddings using supervised learning on a large silver standard sentiment dataset for twitter, which was labeled by means of the emoticons present in the tweets.

3 DATASETS

In this work, we use two kinds of datasets, existing hand labeled datasets (gold standard), as well as the silver standard *wiki suggestion* dataset which is introduced in this work. Table 2 lists the statistics of the used datasets. Since it is a sentence classification task, the train and test datasets comprise of sentences labeled with their respective classes *suggestion* and *non-suggestion*.

3.1 Gold Standard Datasets

As we stated previously, some hand labeled datasets are already available from the previous works. We use the labeled datasets provided by Negi et al. [7] [6], who have performed a detailed annotation study for suggestion mining, and have set guidelines for hand labeling of datasets. In their datasets, only those sentences where suggestions are explicitly expressed are labeled as suggestions while all other sentences are labeled as non-suggestions, including the

ones from which a suggestion can be inferred. For example, ‘I recommend the cup cakes at the bakery next door’, and, ‘The cup cakes from the bakery next door were delicious’, are explicit and implicit forms of a suggestion respectively. The domains which are covered by these datasets include: hotel reviews, travel discussion forums, suggestion forums for software improvement, and twitter. We use all these datasets except the twitter dataset, which we consider out of scope for the current work.

We perform both in-domain as well as cross-domain training of the classifiers in order to evaluate the models for a domain independent training scenario. We employ the hotel review dataset as the gold standard training dataset in all the experiments. For in-domain evaluation of the classifier, we need a hotel domain test dataset. Since the hotel dataset is highly imbalanced with a much smaller number of suggestions than the non-suggestions, further splitting it into train and test samples would result in even lower number of distinct suggestion instances in the training sample. Therefore, we prepare a test dataset of hotel reviews obtained from the same larger sentiment analysis dataset [13] from where hotel training dataset reviews were obtained, making sure that a different set of reviews are picked. Same annotation guidelines were adapted, as provided by Negi and Buitelaar [7]. A total of 5023 sentences were annotated by two annotators, out of which 400 were labeled as suggestions by both the annotators. We use a balanced subset of this dataset by only keeping 400 randomly selected non-suggestions, and removing the rest.

For the cross domain evaluation of a classifier trained on the hotel review dataset, travel discussion and software improvement datasets are used as test datasets. In this case also, we use the balanced subsets of the original versions of these datasets by removing some negative instances.

Table 2 provides an overview of all the used datasets. It has been observed from the previous works that different domains tend to have different proportions of suggestions. In order to evaluate a single model for both domain specific and domain independent training, we want to avoid any bias of domain specific class distribution in the trained model. Therefore we use balanced samples in both training and evaluation datasets. We balance out the training dataset by oversampling the minority class.

3.2 Wiki Suggestion Dataset: A Silver Standard

Due to the lack of large training datasets and a sparse distribution of suggestion instances in the existing datasets, an obvious intuition is to look for existing sources of text on the web where suggestions are already explicitly marked.

One such potential source for obtaining suggestion expressing sentences is web based suggestion forums for different products and services. An example of such suggestion forum is, a dedicated forum to convey the ideas for improvements in Microsoft Office 365 which is hosted by the *uservoice* platform¹. The *uservoice* platform provides customer feedback management service to brands. A first impression about this suggestion forum is that it will mostly contain positive class instances for a suggestion dataset and can be directly used as a silver standard without needing any further human annotation. However, we observe that it is not the case, since

¹<https://office365.uservoice.com/forums/273493-office-365-admin>

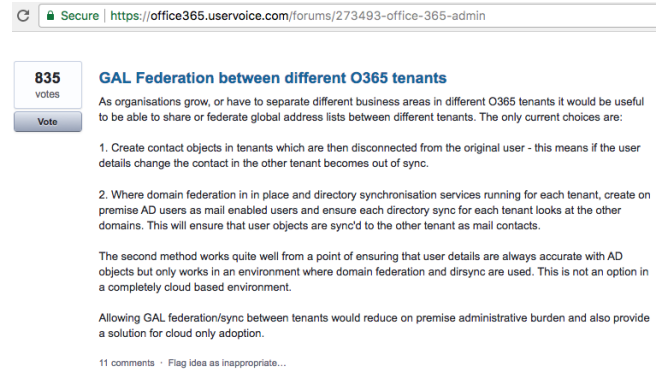


Figure 1: Example of a post from the dedicated forums for users to provide suggestions for new features in brands

text from these posts also contain many elaborative and conversational sentences in addition to the suggestions, which comprise of more than 50% of the sentences in the posts. Therefore such datasets would also require further human annotation, although human labeling would result in training datasets with much higher percentage of suggestions compared to many other sources.

WikiHow is an online wiki-style community consisting of an extensive database of how-to guides, which spans across a large variety of topics. WikiHow articles are open domain, and range over factual and non-subjective topics like *Clean Shoe Insoles*, to opinionated topics like *Resolve Conflict Effectively*. Each article always comprises of a *Steps* section which lists the main steps of doing a certain thing (topic of the page). However, what interests us most is that there are optional sections like *Tips*, *Warnings*, and *Community Q and A*, in addition to the steps section. The tips and warning sections contain short and to-the-point list of suggestions and advice about the topic of the article, and are very much alike the sentences which qualify for the positive instances of a suggestion mining training dataset. In order to obtain an equally large number of negative instances i.e. non-suggestions, we randomly choose equal number of sentences from Wikipedia. Wikipedia mainly contains factual descriptions, and is therefore very less likely to contain expressions of suggestions.

We still consider this dataset as a silver standard, due to three main reasons. Firstly, we do not perform a manual check on each sentence, and there are chances that the labels may not fully adhere to the annotation guidelines used for the gold standard datasets. Secondly, the negative sample from Wikipedia is much cleaner as compared to the ambiguous negative instances in the gold standard datasets. Thirdly, the suggestion expressions on wikiHow are likely to be less varied than the gold standard datasets.

Pre-processing: Preprocessing was only performed on the wikiHow data. We extracted all the items under the tips and warnings sections. These items can sometimes be longer than one sentence. We observe that the main tip is mostly provided in the first sentence, and the following sentences contain further details and explanations, which may/maynot carry an explicit expression of suggestion. Therefore, we perform automatic sentence splitting of each item

Dataset	Source/ Domain	Sentences	Vocabulary	Tokens
Wiki suggestion	Wikihow	675,851	156,768	10,739,320
	Wikipedia	675,851	548,505	20,863,958
Hotel Train	Reviews for different hotels across different cities	Total: 7,534 Suggestions: 448	9,349	1,24,721
Hotel Test (new)	Reviews for different hotels across different cities	Total: 800 Suggestions: 400	2,624	15,390
Travel	Travel discussion threads	Total: 2,618 Suggestions: 1,308	5,724	48,734
Software	Userve feedback forum for <i>Windows</i> operating system, and <i>Feedly</i> mobile app	Total: 2,854 Suggestions: 1,427	6,717	66,717

Table 2: Details of the used datasets

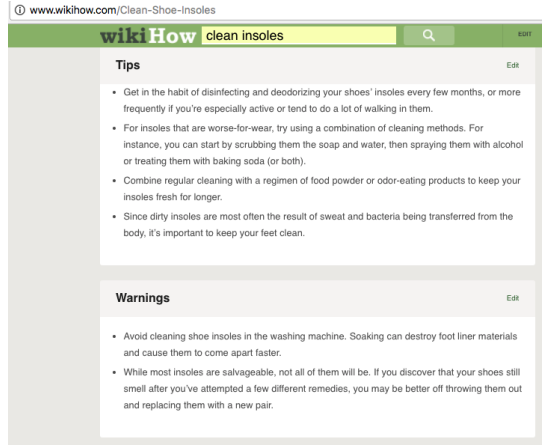


Figure 2: Tips and Warning Sections of the Wikihow article on *How to Clean Shoe Insoles*

and only retain the first sentence as a suggestion instance. The sentence splitting is performed using NLTK’s [4] sentence tokenizer. Before performing the sentence split, we use regular expressions to remove URLs, any further lists enclosed within an item of the main list of tips, and any content within brackets. This results in about 675,851 sentences from wikihow, and thus the final wiki suggestion dataset contains about 1.3 million sentences.

4 METHODOLOGY

We propose two approaches for using the new wiki suggestion dataset for training the classifiers. As mentioned in the Related Work section, Long Short Term Memory neural networks are used for classification in both the approaches. Each approach comprises of different variations.

4.1 Approach 1: Training data for the classifier

In this approach, we induce distant supervision through the wiki dataset by using it as a gold standard training dataset. Due to the relatively smaller size of wiki dataset and hotel training

dataset, as well as our objective to perform cross domain evaluation of models, we use the full vocabulary and initial weights of pre-trained word embeddings from the GloVe² word vectors [8], which are trained on a very large English language corpora built by combining Wikipedia and the Gigaword corpus³ (6 billion tokens). The word vector weights from GloVe are updated each time we train the classifiers; therefore, the words which are not present in the training corpus retain the weights from GloVe. Our implementation of LSTM based classifier also saves the weights of full GloVe vocabulary when a model is saved at the end of the training process. Therefore, if the vocabulary of train and test sets do not match, the model uses the original weights from GloVe to produce the input test data representations for a trained model. We use the following dimensions of the pre-trained GloVe vectors: 200, 100, and 50. This approach is evaluated with the following variations in the training data:

- **Wiki dataset:** Only wiki suggestion dataset is used for training the model.
- **Wiki and hotel dataset:** Training is performed in two passes. The classifier is first pre-trained on the entire wiki dataset. The wiki trained model is then fine-tuned on the hotel train dataset.
- **Semantic subsample of wiki dataset:** A subsample of wiki dataset which is semantically related to the test dataset is used for training. The subset is automatically identified using the following steps:
 - (1) Obtain a bag of all the unique words of the test set.
 - (2) Obtain a vector of these bag of words by performing an addition of the corresponding word vectors (GloVe).
 - (3) Obtain a vector for each sentence in the wiki dataset by using the same method as above.
 - (4) Calculate the similarity score of the test set vector with each of the wiki sentence vector.
 - (5) Choose the top 1% similar sentences from both the classes of the wiki dataset.

²<https://nlp.stanford.edu/projects/glove/>

³<https://catalog.ldc.upenn.edu/ldc2011t07>

This results in a balanced subset of about 13,500 sentences. The 100 dimensional pre-trained embeddings are used to obtain this semantic subsample. Table 3 shows the most sentences with the highest similarity score to the respective test datasets.

- **Wiki subsample and hotel dataset:** Same as the variation #2, except that the semantic subsample of wiki dataset is employed in place of the full wiki dataset.

Baseline: The baseline for this approach is to use a gold standard dataset for training, which is the hotel train dataset in this work.

4.2 Approach 2: Training data for word embeddings

In this approach, wiki dataset is used to induce distant supervision in the classifier model through the pre-trained word embeddings. Instead of using the standard pre-trained GloVe word vectors, we first train word embeddings using the wiki suggestion dataset, and then replace the pre-trained GloVe vectors in the previous approach with these vectors. The neural network classifier is then trained using only the gold standard dataset. Following are the different methods we employed to learn embeddings from the wiki dataset:

- (1) **Wiki Suggestion Embeddings (WiSE):** We train a classifier using the full wiki suggestion dataset, without using any pre-trained word embeddings for the initial weights. Word weights are then extracted from the embedding layer at the end of the training. The classification architecture remains same as that of the approach 1. We refer to these embeddings as the Wiki Suggestion Embeddings (WiSE). This method is different from the state of the art word co-occurrence based or neighbouring word prediction based learning (GloVe, Word2vec respectively) of embeddings, since these embeddings are learned in a supervised manner by means of a class prediction objective.

Specific syntactical and grammatical properties are strong features for suggestions, and suggestions across domains tend to exhibit similar syntactic properties [6]. Earlier works on suggestion mining also show that POS n-gram and POS pattern based features are useful with SVM classifiers [7]. Therefore, using word vectors which strongly encode syntactic properties may be a promising direction for open domain training of suggestion mining classifiers. It has been shown earlier that the smaller dimensions of SOTA word embeddings like GloVe tend to capture syntactic properties of the words [9] and perform well for syntactic analogy tasks. Table 4 also indicate that smaller dimensions of GloVe are performing better than 200 dimensions in some of the experiments. We learn embeddings for part of speech tags, which are not likely to carry any semantic information as an alternative to the standard 50 dimensional word embeddings which carry some semantic information as well.

We investigate two variants of the WiSE embeddings:

- **WiSE_w:** These are embeddings for words. The vocabulary comprises of words which are present in the wiki suggestion dataset, and is therefore smaller than that of

the GloVe vectors. These embeddings are trained for 100 dimensions, since the baseline model also uses 100 dimensions. This variant comprises of about 339,000 words.

- **WiSE_p:** These are the embeddings for Part of Speech tags. These embeddings are very light weight, with 50 dimensions and a vocabulary of size 34. These embeddings are learned by replacing all the words in the wiki corpus with their POS tags.

- (2) **Glove-Wiki_p:** These are the POS embeddings learned using the GloVe algorithm, but trained on the wiki suggestion dataset.

- (3) **Concatenation of Word and POS embeddings:** These are the embeddings obtained by concatenating the best performing word embeddings with the best performing POS embeddings.

Baseline: The baseline for models trained using wiki dataset embeddings is the model which performed best when pre-trained GloVe vectors were used with the gold standard dataset, i.e. the model trained using 100 dimensional pre-trained GloVe (2nd model of the 1st row in Table 4).

4.3 Network Architecture and Hyper-parameters

We evaluated some commonly used configurations and hyper-parameter values for short text classification. These configurations were evaluated on the baseline approach, i.e. using hotel train dataset with all three dimensions of the GloVe vectors. We chose the configurations which perform best for at-least two out of three test datasets.

The final architecture (Figure 3) comprises of the input layer of LSTM units, and a softmax activation is applied to the outermost dense layer of size 2. Two lstm hidden layers of size 50 and 20 units respectively. Tanh activation is used at each layer, L2 regulariser is used with the first three LSTM layers, Adam is used for optimisation, and the training was performed with 5 epochs. Keras (theano backend), Gensim, and Scikit learn libraries are primarily used for the implementation.

5 RESULTS AND ANALYSIS

Tables 4, 5 show results for approaches 1 and 2 respectively. Precision, Recall, and F1 score of the suggestion class (positive class) on different test datasets are reported. We compare the performance of different models on the basis of F1 score.

For approach 1, using the full wiki dataset as the training dataset, which is an open domain training dataset, outperforms all other variants of training datasets for all three test datasets. However, the performance improvement over baseline for the domain whose training dataset is available (hotel) is small, as compared to the other domains where domain specific training was not performed. For approach 2, results demonstrate that the use of pre-trained POS embeddings yields better results as compared to the baseline of

	Hotel	Travel	Software
wikiHow	Even though most of the Universal Orlando Parks have a few areas where kids can explore and run out all their excitement, this may sometimes not be enough.	If you reacted a certain way because they did a certain thing, you'll want to bring it up tactfully.	At this time, Keek only allows users to delete their accounts from the Keek.com website, however, this feature will soon be available for Keek apps on iOS, Android, Blackberry 10, and Windows 8 devices.
Wikipedia	Several important people of the city of Clermont had asked me to let them know when I would make the ascent.	So I thought,'ok, I'll carry on doing this for a bit and the next thing you know that's how I make my living these days'.	The HD's user interface, he noted, was the first such Microsoft product to rely on text rather than icons, and it would form the basis for Windows Phone, Windows 8, Xbox and all of the company's web-based services.

Table 3: Examples of sentences from the two classes in the wiki dataset with the highest similarity score with the different test datasets.

Train data	Embedding	Hotel Test			Travel			Software		
		P	R	F1	P	R	F1	P	R	F1
Hotel (Baseline)	Glove 200	0.99	0.59	0.74	0.78	0.26	0.40	0.70	0.28	0.40
	Glove 100	0.95	0.71	<u>0.81</u>	0.71	0.35	<u>0.47</u>	0.64	0.49	<u>0.55</u>
	Glove 50	0.97	0.57	0.72	0.72	0.28	0.40	0.69	0.39	0.50
Wiki	Glove 200	0.75	0.90	0.82	0.57	0.70	0.63	0.55	0.75	0.64
	Glove 100	0.76	0.88	0.82	0.56	0.69	0.62	0.55	0.76	0.64
	Glove 50	0.76	0.88	0.82	0.56	0.70	0.62	0.56	0.79	0.66
Wiki + Hotel	Glove 200	1.00	0.29	0.45	0.78	0.06	0.12	0.78	0.07	0.13
	Glove 100	1.00	0.38	<u>0.55</u>	0.83	0.14	0.23	0.71	0.21	0.32
	Glove 50	1.00	0.38	<u>0.55</u>	0.81	0.18	<u>0.29</u>	0.73	0.33	<u>0.45</u>
Subsampling Wiki Dataset										
Wiki subsample	Glove 200	0.25	0.49	<u>0.33</u>	0.61	0.46	<u>0.53</u>	0.61	0.46	0.53
	Glove 100	0.99	0.19	0.32	0.70	0.14	0.23	0.52	0.74	<u>0.61</u>
	Glove 50	0.25	0.49	<u>0.33</u>	0.74	0.22	0.33	0.52	0.68	0.59
Wiki subsample + Hotel	Glove 200	0.51	0.86	0.63	0.50	0.86	0.63	0.52	0.75	0.62
	Glove 100	0.99	0.59	<u>0.74</u>	0.82	0.14	0.24	0.64	0.12	0.21
	Glove 50	0.50	1.00	0.67	0.50	1.00	<u>0.67</u>	0.50	1.00	<u>0.67</u>

Table 4: Results from different dataset settings using gold standard hotel and Wiki Suggestion dataset to directly train a classifier. Precision, Recall, and F measure are provided for the positive class, i.e. the suggestion class.

Embedding	Dimensions, Vocabulary	Hotel Test			Travel			Software		
		P	R	F1	P	R	F1	P	R	F1
Word Embeddings										
Pre-trained Glove (Baseline)	100; 400,000	0.95	0.71	<u>0.81</u>	0.71	0.35	0.47	0.64	0.49	<u>0.55</u>
WiSE _w	100; 339,240	0.26	0.51	0.34	0.25	0.50	0.33	0.25	0.50	0.33
Part of Speech Embeddings										
Glove (Baseline)	50; 400,000	0.97	0.57	0.72	0.72	0.28	0.40	0.69	0.39	0.50
Glove-Wiki _p	50, 34	0.80	0.83	0.82	0.57	0.71	0.63	0.57	0.90	0.70
WiSE _p	50, 34	0.82	0.84	<u>0.83</u>	0.57	0.69	0.62	0.58	0.90	<u>0.71</u>
Word + Part of Speech Embeddings										
Glove + WiSE _p Concat	150; 400,000 * 34	0.90	0.80	0.85	0.62	0.57	0.59	0.61	0.81	0.70

Table 5: Classification results of using different embeddings with hotel data as the training dataset. Precision, Recall, and F1 score for the suggestion class are presented.

using pre-trained GloVe embeddings. Based on the current experiments, approach 2 appears to be a better method to employ wiki suggestion dataset for suggestion mining. The results shown in

Table 5 are very interesting considering the fact that the training vocabulary for approach 2 (only POS tags) is much smaller than

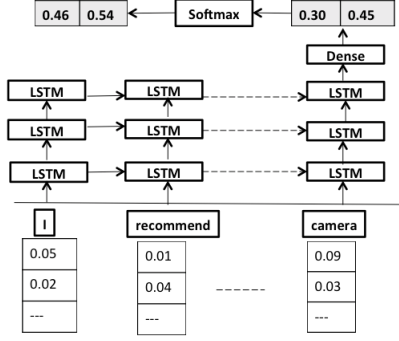


Figure 3: Architecture for the LSTM based neural network classifier

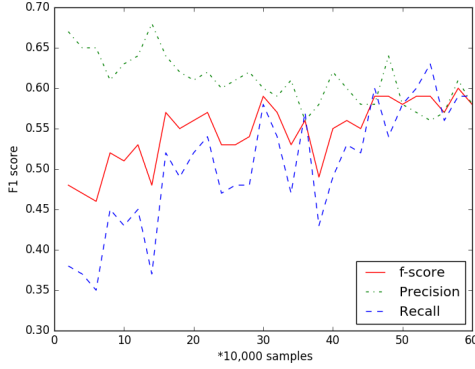


Figure 4: Training sample size of wiki dataset vs. precision, recall and F1 score on the travel dataset

the standard word embeddings. However, the performances of POS embeddings learned using WiSE and GloVe methods are very close.

5.1 Qualitative and Quantitative Analysis of Wiki Dataset

Approach 1: Using full wiki dataset yielded best results, while using its subsamples significantly reduced the performance. Travel test data showed the largest improvement over baseline on using wiki dataset as training dataset. Figure 4 shows the change in precision, recall, and F1 score with the increasing size of wiki training dataset in approach 1, which hints that the size of the wiki dataset is the reason for its better performance over the much smaller gold standard dataset, since the results show improvements with large increases in its sample size. However, using the smaller samples of wiki dataset significantly reduced the performance despite of these samples being semantically close to the test datasets. This can also be observed in Figure 4 which shows the steep variations in performance within samples of comparable size.

Wiki dataset for POS embeddings: Software test data showed the largest improvement in F1 score on using $WiSE_p$ embeddings. Figure 5 shows the change in precision, recall, and F1 score of the

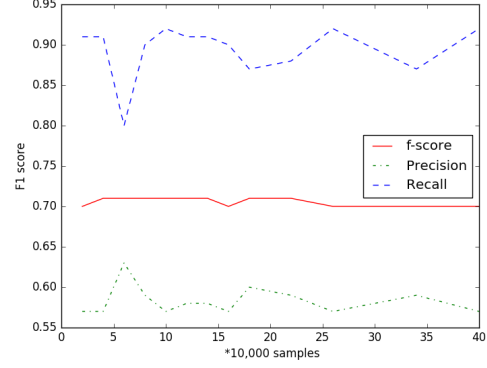


Figure 5: Training sample size for WiSE vs. precision, recall and F1 score on the software dataset

Test Data	P	R	F
Hotel	0.58	0.50	0.53
Travel	0.52	0.55	0.53
Software	0.53	0.59	0.56

Table 6: Classification results when POS version of wiki dataset is directly used to learn the classification model, which also includes learning $WiSE_p$.

Dataset	Vocab	Overlapping Vocabulary	
		wikiHow	Wikipedia
Hotel	2624	289	219
Travel	5724	838	620
Software	6717	1643	1557

Table 7: Overlapping vocabulary of test datasets with the wiki dataset.

software test dataset when the POS embedding are learned on the different sized subsamples of wiki dataset in approach 2, which shows that the large size of dataset doesn't contribute much in this case. This is in line with our observation that the style of suggestion expressions do not vary much in this dataset. Table 6 show the results when we use only the POS version of wiki dataset to train the classifier in approach 2, but the results are much lower compared to using the gold standard (Table 5). This hints that the model which uses POS embeddings can model syntactic variations in suggestions with a much smaller training dataset as compared to using word embeddings.

Wiki dataset for word embeddings: In approach 2, $WiSE_w$ performed poorly as compared to all other embeddings. A possible reason could be that a large portion of test dataset vocabulary is not present in the wiki dataset (Table 7), therefore the vocabulary of the $WiSE_w$ embeddings is insufficient for learning task specific word embeddings.

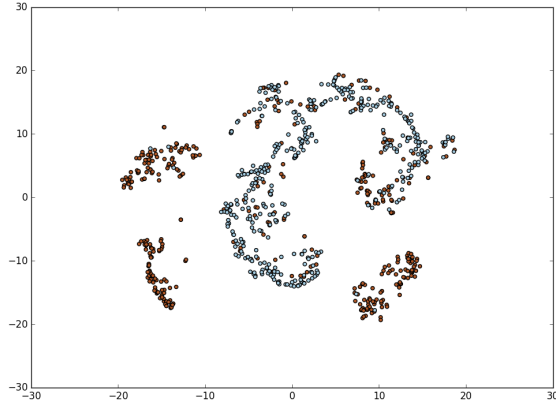


Figure 6: Sentence representations of hotel test dataset obtained from the last layer of the model trained using 100 dimensional glove embeddings and hotel train dataset.

5.2 Word vs POS Representations

Figures 6 and 7 plot sentence vectors of hotel test instances on a 2-dimensional plot using the t-sne method of dimensionality reduction [12]. The sentence vectors are obtained from the last LSTM layer of the models trained on the hotel train dataset, using pre-trained GloVe vs. WiSE_p embeddings. Red and blue dots show the original labels for positive and negative instances respectively. We can observe high variability within the representations of samples of the positive class (red) as shown by the number of clusters of positive instances, in the GloVe embeddings based sentence representations (figure 6). This may indicate the variations among the topics of suggestions. In contrast to this, WiSE_p (POS embeddings) based sentence representations result in 2 major clusters for the positive instances (figure 7), indicating lesser syntactic variations among the suggestion instances. This may lead to an easier decision boundary in comparison to the former, and can explain the results described in the previous sub-section.

6 CONCLUSION

Our experiments have revealed several interesting observations. Most interesting results are the high performance when using pre-trained POS embeddings with part of speech versions of the datasets, which is an extremely lightweight model. This also provides a ground for evaluating POS embeddings based approach on other sentence level tasks.

The two proposed approaches to use wikihow dataset perform better than the respective baselines, the POS embedding approach in particular outperforms all other variations (across both the approaches). This proves that syntax and grammar are predominant features in determining the suggestion class. However, it is also evident that in order to achieve higher classification accuracy, domain data would still be required. Last but not least, to the best of our knowledge this is the first attempt to use wikiHow for a text classification task.

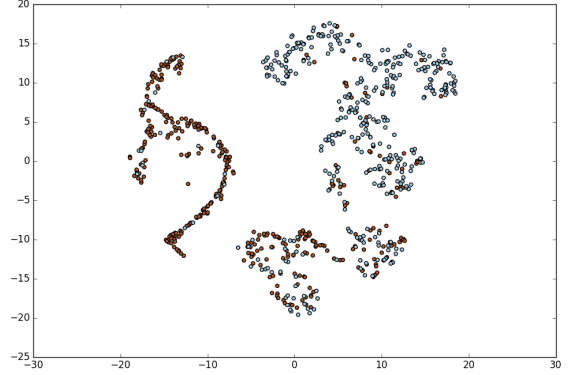


Figure 7: Sentence representations of hotel test dataset obtained from the last layer of the model which was trained using WiSE_p embeddings and hotel train dataset.

In future we would like to investigate different neural network architectures to compose the word and POS embeddings against the simple concatenation method used in this work.

REFERENCES

- [1] Marco Baroni, Georgiana Dinu, and Germán Kruszewski. 2014. Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Baltimore, Maryland, 238–247. <http://www.aclweb.org/anthology/P14-1023>
- [2] C. Brun and C. Hagege. 2013. Suggestion Mining: Detecting Suggestions for Improvements in Users Comments. In *Proceedings of 14th International Conference on Intelligent Text Processing and Computational Linguistics*.
- [3] Li Dong, Furu Wei, Yajuan Duan, Xiaohua Liu, Ming Zhou, and Ke Xu. 2013. The Automated Acquisition of Suggestions from Tweets.. In *AAAI*, Marie desJardins and Michael L. Littman (Eds.). AAAI Press.
- [4] Edward Loper and Steven Bird. 2002. NLTK: The Natural Language Toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1 (ETMTNLP '02)*. Association for Computational Linguistics, Stroudsburg, PA, USA, 63–70. <https://doi.org/10.3115/1118108.1118117>
- [5] Samaneh Moghaddam. 2015. Beyond Sentiment Analysis: Mining Defects and Improvements from Customer Feedback. In *Advances in Information Retrieval*, Allan Hanbury, Gabriella Kazai, Andreas Rauber, and Norbert Fuhr (Eds.). Lecture Notes in Computer Science, Vol. 9022. Springer International Publishing, 400–410. https://doi.org/10.1007/978-3-319-16354-3_44
- [6] Sapna Negi, Kartik Asooja, Shubham Mehrotra, and Paul Buitelaar. 2016. A Study of Suggestions in Opinionated Texts and their Automatic Detection. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*. Association for Computational Linguistics, Berlin, Germany, 170–178. <http://anthology.aclweb.org/S16-2022>
- [7] Sapna Negi and Paul Buitelaar. 2015. Towards the Extraction of Customer-to-Customer Suggestions from Reviews. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 2159–2167. <http://aclweb.org/anthology/D15-1258>
- [8] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543. <http://www.aclweb.org/anthology/D14-1162>
- [9] Barbara Plank, Anders Søgaard, and Yoav Goldberg. 2016. Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss. *arXiv preprint arXiv:1604.05529* (2016).
- [10] J Ramanand, Krishna Bhavsar, and Niranjan Pedaneekar. 2010. Wishful Thinking - Finding suggestions and 'buy' wishes from product reviews. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*. Association for Computational Linguistics, Los Angeles, CA, 54–61. <http://www.aclweb.org/anthology/W10-0207>

- [11] 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)*. Association for Computational Linguistics, Baltimore, Maryland, 1555–1565. <http://www.aclweb.org/anthology/P14-1146>
- [12] Laurens Van Der Maaten. 2014. Accelerating t-SNE using tree-based algorithms. *Journal of machine learning research* 15, 1 (2014), 3221–3245.
- [13] Henning Wachsmuth, Martin Trenkmann, Benno Stein, Gregor Engels, and Tsvetomira Palakarska. 2014. A Review Corpus for Argumentation Analysis. In *Proceedings of the 15th International Conference on Computational Linguistics and Intelligent Text Processing (LNCS)*, Vol. 8404. Springer, Kathmandu, Nepal, 115–127.
- [14] Alfian Farizki Wicaksono and Sung-Hyon Myaeng. 2013. Automatic extraction of advice-revealing sentences for advice mining from online forums. In *K-CAP*. ACM, 97–104.