# Literature Review for the Final Project - Stanford NLU Course

Author: Fiodar Ryzhykau

Stanford University

XCS224U – Online Professional Course

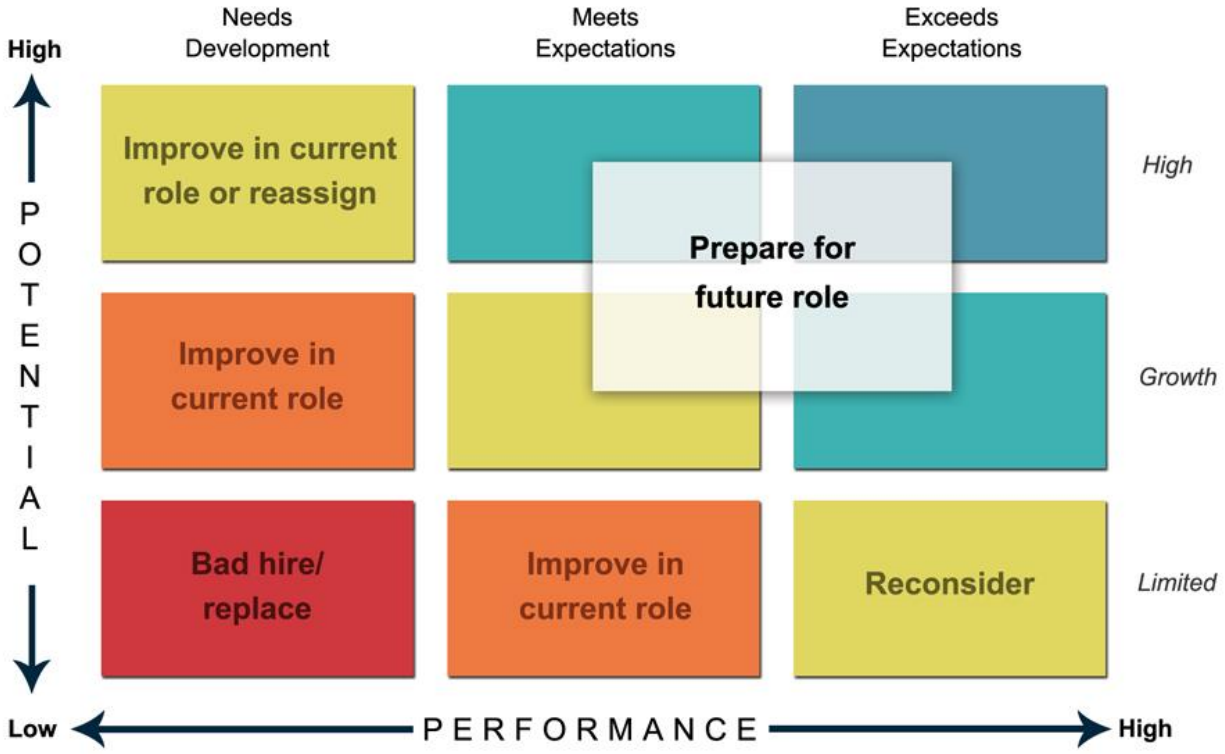
October 4th 2020

# Literature Review: Using NLU methods for automated employee review analysis

For my Final Project in the Stanford NLU course I've selected a topic that is related to the employee assessment in modern companies. In my current company, every quarter (and annually) managers should provide feedback to their team-members and grade their Performance and Potential in order to objectify the value of that employee to the company.

We use the "9-box model" (or 9 grid model), which is also called the "Performance and Potential Model" and has been developed by McKinsey. This approach is used in many companies as a method to identify, support and promote the talent.

In short, these "9-boxes" look the following way:




Feedback (review) that is being submitted along the 9-box grade should explicitly describe all the strong and weak points of the employee, as well as sentiment to the contribution and progress for the given quarter. Submitted data basically represents a mapping/summarization of the feedback to one of the 9 categories.

Considering the learnings from the NLU course I'm targeting to apply the machine learning technics to automate 9-box grading based on the manager's feedback.

Definition of the problem relate to 2 potential directions in the NLU space: Sentiment Analysis and Article Summarization.

One of the challenges for the given task is the lack of data available on the web (most likely due to its PII characteristics). The dataset that will be collected with the help of Amazon MTurk crowd will be limited by just several hundreds of examples, and thus it's important to research the approaches applicable to such limitation.

Another area of focus for literature review is Transfer Learning with existing powerful models like BERT.

# Outline of Existing Works

The following resources have been reviewed:

| Category | Paper |
|---|---|
| Sentiment Analysis with classic SVM-based models | - **UDLAP: Sentiment Analysis Using a Graph Based Representation** (2015) by Esteban Castillo, Ofelia Cervantes, Darnes Vilarino, David Baez and Alfredo Sanchez<br>- **Sentiment Analysis of Twitter Data** (2011) by Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau |
| Using BERT for Sentiment Analysis | - **Aspect-Based Sentiment Analysis Using BERT** (2019) by Mickel Hoang, Oskar Alija Bihorac, Jacobo Rouces<br>- **FinBERT: Financial Sentiment Analysis with Pre-trained Language Models** (2019) by Dogu Tan Araci |
| Using BERT for Text Summarization | - **Text Summarization with Pretrained Encoders** (2019) by Yang Liu, Mirella Lapata |

# Overview

## Paper 1: UDLAP: Sentiment Analysis Using a Graph Based Representation

Explains how Graph-based representation of the text/messages perform comparing to the more classic (linear, feature-based) on Message Polarity Classification (SemEval-2015 Task 10, Subtask B) which is similar to my initial problem statement if grading axes are considered separately (Low/Medium/High for Performance and Potential): ”Given a message, classify whether the message is of positive, negative, or neutral sentiment. For messages conveying both a positive and a negative sentiment, whichever is the stronger sentiment should be chosen”. Results show that such models can outperform the baseline for avg ~1.5 times in F1 score (F1 = [F1pos + F1neg]/2) , however LiveJournal and SMS datasets scored lower than Twitter, potentially due to the size of the messages or a slang usage.

**Dataset:** Unbalanced, 7.5k tweets for training. **Data pre-processing:** elimination of punctuation symbols and all the elements that are not part of the ASCII encoding. **Tokenization:** not mentioned.

## Paper 2: Sentiment Analysis of Twitter Data

Covers details for similar task as Paper 1, but provides more details about data pre-processing, tokenization and leveraging synonym dictionaries like WordNet. SVM with tree kernel outperformed senti-feature-based and the state-of-the-art unigram baseline model, and

showed similar F1 results for 3 classes (Positive/Neutral/Negative) as for the 2 class scenario. Combination of tree kernel + senti-feature model produced the best result (F1: ~60%) and outperformed tree kernel model. For senti-feature model - polarity of words and their parts-of-speech tags provided the highest gain, which drives to conclusion that word dictionaries like DAL (Dictionary of Affect in Language) could be helpful for this task.

**Dataset:** 5.1k tweets for training with balanced distribution of 1.7k per class. Tweets limited with 140 symbols. **Tokenization:** Stanford tokenizer (Klein and Manning, 2003) + stop word dictionary. **Data-preprocessing:** emoticon dictionary and an acronym dictionary.

### Paper 3: Aspect-Based Sentiment Analysis Using BERT

Shows the potential of using the contextual word representations from the pre-trained language model BERT, in order to solve out-of-domain ABSA (Aspect-based sentiment analysis) and outperform state-of-the-art results on SemEval2015 (task 12, subtask 2) and SemEval2016 (task 5). It proposes a combined model, which uses only one sentence pair classifier model from BERT to solve both aspect classification and sentiment classification simultaneously. In out-of-scope scenarios, the classifiers which have been trained on sentence-level datasets outperform the classifiers which have been trained on the text-level datasets. Also, models trained on combined datasets (Restaurant + Laptop) outperformed the ones on trained on single. F1 scores for single and combined models ranged from 82% to 90%.

**Dataset:** Restaurant and Laptop reviews for training (size of 2000/334 and 2500/395 sentences/texts respectively), and both of them together with Hotel reviews (out-of-domain) for test. **Tokenization:** wordpiece tokenization (Wu et al., 2016) + 2 specialized tokens: classifier token [CLS], which is added to the beginning of the set; and separation token [SEP], which marks the end of a sentence.

### Paper 4: FinBERT: Financial Sentiment Analysis with Pre-trained Language Models

This work describes application of BERT for finance and experimentation with further pre-training on a domain-specific corpus. In addition to BERT, ELMo and ULMFit language models were trained for comparison purposes. ULMFit, further pre-trained on a financial corpus, beat the previous state-of-the art for the classification task, only to a smaller degree than BERT. These results show the effectiveness of pre-trained language models for a down-stream task such as sentiment analysis especially with a small labeled dataset. The complete dataset included more than 3000 examples, but FinBERT was able to surpass the previous state-of-the art even with a training set as small as 500 examples.

**Dataset:**. The main sentiment analysis dataset used in this paper is Financial PhraseBank5 from Malo et al. 2014, which consists of 4845 English sentences selected randomly from financial news found on LexisNexis database. Labeled according to how annotators think the information in the sentence might affect the mentioned company stock price. 20% of all sentences were set as test and 20% of the remaining as validation set. In the end, train set included 3101 examples.

The chart below provides a good view on the correlation of model performance with training set size:

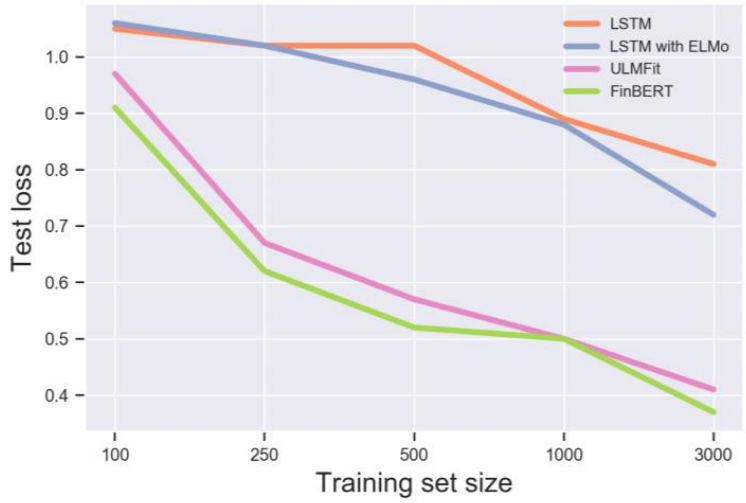



**Figure 2: Test loss different training set sizes**

## Paper 5: Text Summarization with Pretrained Encoders

In this paper, authors showcased how pretrained BERT can be usefully applied in text summarization. Extractive summarization systems create a summary by identifying (and subsequently concatenating) the most important sentences in a document. Abstractive summarization conceptualizes the task as a sequence-to-sequence problem, where an encoder maps a sequence of tokens in the source document to a sequence of continuous representations, and a decoder then generates the target summary token-by-token, in an auto-regressive manner. Experimental results across three datasets showed that model achieved state-of-the-art results across the board under automatic (ROUGE-1, ROUGE-2, ROGUE-L) and human-based evaluation protocols.

**Datasets:** #docs (train/val/test): CNN 90,266/1,220/1,093; DailyMail 196,961/12,148/10,397; NYT 96,834/4,000/3,452; XSum 204,045/11,332/11,334

# Comparative Review and Contrasting Opinions

Considering the intuitively selected direction of research, given papers were chosen with the goal to understand the approaches people take in order to solve Sentiment Analysis and Summarization tasks in NLU field. It's becoming clear that in the recent few years, pretrained language models like BERT have started to quickly drive the state-of-the-art performance.

Focus of the papers dated 2011 and 2015 (**"Sentiment Analysis of Twitter Data"** and **"UDLAP: Sentiment Analysis Using a Graph Based Representation"** respectively) was set on finding better word/text representations with high degree of data pre-processing. However in most of the cases the model performance boost was marginal (considering the state-of-the-art baselines of that time).

Papers dated 2019+ (**"Aspect-Based Sentiment Analysis Using BERT"**; **"FinBERT: Financial Sentiment Analysis with Pre-trained Language Models"**) have shifted focus on construction of encoders/decoders and selection of the proper set of layers and their configurations for the NN-based pretrained models. Such context-based models have also shown a good transfer-learning capability for cross-domain data.

Another important conclusion from **"FinBERT: Financial Sentiment Analysis with Pre-trained Language Models"** paper, states that deep learning techniques for NLP, that previously were known to be "data-hungry" should apparently no longer be considered as such.

**"Text Summarization with Pretrained Encoders"** paper provided a good overview of the text summarization approaches, leveraging pretrained BERT encoders. The aim of the summarization is to condense a document into a shorter version while preserving most of its meaning, which slightly differ from my Final Project task definition. One important finding that can still be considered for the long employee reviews – extractive summarization models have shown that most important document content from the News dataset were located within approximately 5 first sentences (according Human-validated criteria of Informativeness, Fluency, and Succinctness):

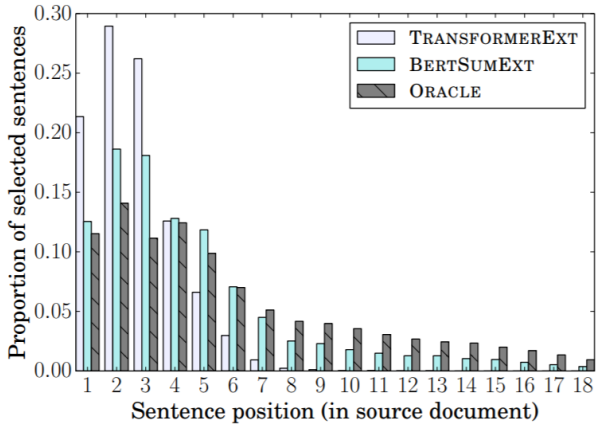



Figure 2: Proportion of extracted sentences according to their position in the original document.

# Conclusions and Future Work

Reviewed papers have significantly helped to clarify the direction and approach for the future work. Based on the recent papers it's getting pretty obvious that transfer-learning method with state of the art language models like BERT overtake the classic approaches in NLP/NLU tasks. In terms of the NLU task definition the 9-box classification looks very similar to the sentiment analysis, grounded to a particular (in my case employee review) domain.

Paper 4 (FinBERT) has shared a valuable view on the dataset size requirements for pretrained language models, which drives to a conclusion that MTurk-collected dataset may be sufficient to achieve decent results.

However it looks like classic SVM-based models are still a bit more straightforward to use, and may require less effort to kickstart. Thus it makes me think that it could be a good candidate for initial baseline.

# References


1. Hello Monday (website), 30 April 2018, **9 box model: adding another dimension to performance reviews**
2. Mickel Hoang, Oskar Alija Bihorac, Jacobo Rouces, Sept-Oct 2019, **Aspect-Based Sentiment Analysis Using BERT**, ACL Anthology ID: W19-6120
3. Dogu Tan Araci, Aug 2019, **FinBERT: Financial Sentiment Analysis with Pre-trained Language Models**, arXiv:1908.10063
4. Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, June 2011, **Sentiment Analysis of Twitter Data**, ACL Anthology ID: W11-0705
5. Esteban Castillo, Ofelia Cervantes, Darnes Vilarino, David Baez and Alfredo Sanchez, June 2015, **UDLAP: Sentiment Analysis Using a Graph Based Representation** , ACL Anthology ID: S15-2093
6. Yang Liu, Mirella Lapata, Nov 2019, **Text Summarization with Pretrained Encoders**, Anthology ID: D19-1387