NOVA IMS Information Management School

MDSAA

Master Degree Program in
Data Science and Advanced Analytics

Predictive model for detecting fake reviews

Exploring the possible enhancements of using word embeddings

Doris Macean

Dissertation

presented as partial requirement for obtaining the Master Degree Program in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

[this page should not be included in the digital version. Its purpose is only for the printed version]

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

PREDICTIVE MODELS FOR DETECTING FAKE REVIEWS VIA WORD EMBEDDINGS

by

Doris Macean

Dissertation report presented as partial requirement for obtaining the Master's degree in Advanced Analytics, with a Specialization in Data Science

Supervisor: Fernando Bação

November 2022

STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

November 30, 2022

ABSTRACT

Fake data contaminates the insights that can be obtained about a product or service and ultimately hurts both businesses and consumers. Being able to correctly identify the truthful reviews will ensure consumers are able to more effectively find products that suit their needs. The following paper aims to develop a predictive model for detecting fake hotel reviews using Natural Language Processing techniques and applying various Machine Learning models. The current research in this area has primarily focused on sentiment analysis and the detection of fake reviews using various text mining methods including bag of words, tokenization, POS tagging and TF-IDF. The research mostly looks at some combination of quantitative and qualitative information. The text component is only analyzed with regards to which words appear in the review, while the semantic relationship is ignored. This research attempts to develop a higher level of performance by implementing pretrained word embeddings during the preprocessing of the text data. The goal is to introduce some context to the text data and see how each model's performance changes. Traditional text mining models were applied to the dataset to provide a benchmark. Subsequently, GloVe, Word2Vec and BERT word embeddings were implemented and the performance of 8 models was reviewed. The analysis shows a somewhat lower performance obtained by the word embeddings. It seems that in texts of short length, the appearance of words is more indicative of a fake review than the semantic meaning of those words.

KEYWORDS

Natural Language Processing; Machine Learning; Text Mining; Sentiment Analysis; Word Embeddings

INDEX

1. INTRODUCTION

User generated content provides valuable insights that can be an important source of information. While creators of web content are biased and provide information in such a way that their business benefits, user generated content is an unbiased, third party's view on some content, service and/or product. At least, that is the expectation of other web content users.

It goes without saying that reviews have a huge impact on a business's future success. The fact that the internet has become so widely used, easily accessible, and unrestricted has led to an incredible growth of user-generated content, and particularly online product and service reviews (Duan et al., 2016; Zhao et al., 2015). With larger product and service availability, consumers have a multitude of options to fulfill their needs, and no real effective way to evaluate them. In a perfect world, an individual could try out all the available options and be able to decide for themselves. Unfortunately, limited resources of time and money make such an approach unfeasible. A consumer must simply make an educated guess with whatever information is available. Without personal experience, a close second is the experience of others.

Online reviews allow individuals to learn from the experiences of others. The information obtained from these reviews drives the decision-making process. Research shows that online reviews have an impact on up to 50% of hotel booking decisions (Duan et al., 2016). It has also been found that "review ratings are a more significant predictor of hotel performance than traditional customer satisfaction surveys". (Woo Gon Kim and Seo Ah Park, 2017)

In addition to impacting the booking of hotel rooms, it also has a large influence on a hotel's profitability. Previous research has found that a "1% increase in a hotel's rating can lead to a 0.54% increase in occupancy and a 1.42% increase in revenue per available room" (Anderson, 2012). There is evidence that shows the amount a client is willing to pay is proportional to the review ratings (Cantallops and Salvi, 2014; Kwok et al., 2017). Understanding the impact that an online review has on the profitability of the business immediately introduces the threat of fraudulent or deceptive reviews.

Another challenge is the lack of labeled data. Manually validating the reliability of opinions reflected in online reviews is challenging. With no quality control or verification that an individual is in fact a customer, users can write anything on the web. "This results in many low-quality reviews, and worse still review spam". (Jindal & Liu, 2008)

While the prediction of ratings and which features are of value to consumers is essential for the service industry, the first step is ensuring that the data is reliable and accurate. With the reliance on reviews in the decision-making process, as well as a lack of regulation in the practice of making reviews, deceptive reviews will continue to be a huge challenge for many businesses that rely on customer assessments. As such, I am focusing my analysis on investigating how machine learning and natural language processing can more effectively detect these fraudulent reviews to ensure more reliable information is driving the consumer's decision-making process.

2. LITERATURE REVIEW

2.1. THEORETICAL BACKGROUND

There is a multitude of evidence showing that consumers rely heavily on reviews when deciding to purchase a product and/or service. Research shows that 80% of consumers will make a different decision due to many negative reviews, and 87% will choose to make a purchase based on positive reviews. (Tang & Cao, 2020). As such, individuals are vulnerable to potential inaccuracies in online reviews.

The literature defines three types of reviews.

Types of reviews:

1. Untruthful reviews: Reviews that are intended to be misleading. These can be positive to promote a business, or alternatively negative reviews intending to damage a business's reputation (i.e., defaming spam). (Jindal & Liu, 2008)

2. Reviews on brand only: These reviews speak to the brand, manufacturer, or sellers of the products. The quality of the products themselves is not included in such a review. While the information is useful in some regards, these reviews are considered spam as they are not about the actual product and can often be biased. (Jindal & Liu, 2008)

3. Non-reviews: There are two types of non-reviews. The first being advertisements and the second being irrelevant texts that don't contain an opinion. For example, these non-reviews could be asking or answering questions about a product. (Jindal & Liu, 2008)

The detection of fake reviews is a classification problem with two classes, fraudulent and truthful. We must handle the various types of spam in different ways. We can use traditional classification methods to identify spam reviews of type 2 and 3. These types of reviews can be easily categorized manually (Jindal & Liu, 2008). Manual labeling untruthful reviews (type 1) by simply reading the reviews is extremely challenging. Spammers can write a fraudulent review that looks exactly like any other truthful review. Keeping that in mind, it makes sense to focus our efforts on finding an automated way to detect these deceptive reviews.

Deceptive opinion spam is a type of review with fictitious opinions, and it is deliberately written to sound authentic. I will focus on this type of review. Even within this category, some are more harmful than others. A deceptive review that provides a positive assessment for a product and/or service that already has a high average rating is not harmful. This review is simply validating the consensus further and will likely not result in any major changes to a client's purchasing decision. A review that falsely provides a negative assessment of a product and/or service that most clients generally like can be detrimental to a business. In general, an opposing negative review will be much more harmful than an opposing positive review. In the table below you can see a breakdown of spam review types. (Jindal & Liu, 2008).

	Positive spam review	Negative spam review
Good quality product	1	**2**
Bad quality product	**3**	4
Average quality product	**5**	**6**

Table 1: Breakdown of spam review types (Jindal & Liu, 2008).

Manufacturers of a product, or alternatively, individuals that have some interest in the product being successful have an incentive to promote the product. As such, we expect positive review spam in regions 1, 3 and 5 are written by such individuals. Reviews in region 1 are situations where a reviewer has a conflict of interest, and their opinion is not completely factual. That said, their review is consistent with the average quality of the product. Competitors have an incentive to harm their competition's sales via negative review spam. These individuals are likely to write the reviews in regions 2, 4 and 6. While opinions reflected in region 4 may be consistent with the true quality of the product, reviewers have malicious intensions. Deceptive reviews in regions 1 and 4 are not so harmful as they are consistent with the average quality of the product. Reviews in regions 2, 3, 5 and 6 differ from the true quality of the product. These are the most detrimental spam reviews, specifically those in sections 2 and 6 which reflect a negative sentiment. Spam detection processes should focus on detecting reviews in these sections. (Jindal & Liu, 2008).

A major challenge here is that these types of deceptive reviews are extremely hard to identify for the average person. This is where text mining and machine learning come in.

2.2. APPLIED METHODOLOGIES

Now that we understand the data types and the information that exists within reviews, let's review different approaches that have been explored in the literature.

As obtaining correctly labeled data is difficult, the most basic model designed was based on looking at three types of duplicates (Jindal & Liu, 2008). In any case, duplicates are considered fake reviews.

1. Duplicates from different userids on the same product.
2. Duplicates from the same userid on different products.
3. Duplicates from different userids on different products

In this field of research, the Support Vector Machine classifier was the most utilized, followed by Naive Bayes, Decision Tree, Random Forest, and Logistic Regression. Neural networks and ensemble methods have also been explored; however, appear to be less common. Finally, others have proposed prediction models based on semi-supervised learning and a combination of textual and behavioral features. Others have suggested methods for "capturing relationships between reviewers, reviews and products". (Fang et Al. 2020)

Regarding the field of application, previous research has mostly focused on the service industry, particularly within the hospitality industry (i.e., restaurant and hotel domains) (Anderson, 2012; Cantallops and Salvi, 2014; Duan et al., 2016; European Commission, 2014; Kwok et al., 2017; Zhao et al., 2015). Some have also applied their research to e-commerce as well. Existing supervised algorithms developed in literature are typically focused on one

domain and depend on domain-specific language (Li et al., 2014). For example, when classifiers are trained on the hospitality domain, the classifiers may assign high weights to words specific to the domain of interest, such as “hotel”, “rooms”, or even the hotel name (Li et al., 2014). It is unclear whether these classifiers will perform well at detecting deception in other domains

The quantitative data (i.e., numerical information such as ratings) appears to be the focus of much of this work. Understandably so, numerical data is easier to quantify by machines. (Duan et al., 2016; Han et al., 2016; Kwok et al., 2017; Liu et al., 2017). “From a total of 67 articles published, across seven major hospitality and tourism journals, between January 2000 and July 2015, 70% employed quantitative methods, 24% employed qualitative methods, and only 4% employed mixed methods” (Kwok et al.,2017). While most of the analysis has focused on the limited information found in quantitative data, actual clients “seem to favor the information-rich, textual components of reviews” (Noone and McGuire (2014)). From the service provider’s perspective, qualitative reviews “can potentially yield insights not indicated in the ratings for how to improve operations and better meet customer expectations” (Han et al., 2016).

There have been some developments within the area of fake news as well. There appears to be a somewhat higher focus on fake news as the impacts of deceptive fake news are far more meaningful. “Fake news is the greatest threat to our so-called freedom of media, apart from distortion and corrupting ideologies, it has also led to tangible consequences, like cybercrime, phishing, cyber-attacks and the list goes on.”(Agarwal et al., 2020). In addition, news tends to have a lot more text than reviews and as such, there are more features with which to train a model. As we know, more data generally leads to more accurate models. GloVe word embeddings have been used to preprocess text into numerical tokens within a vector space where the semantic meaning was encoded into the features (Agarwal et al., 2020). “GloVe embedding was found to be extremely useful as it provided each word a vector projection which was manipulated by its relation, similarities, dissimilarities with other words in the vocabulary, hence complementing the training process in a much more significant way than what a traditional method of Bag of Words would have done.”(Agarwal et al., 2020)

There has also been some research done using Transfer Learning to detect spam emails. (Tida and Hsu, 2022). Transfer learning can be defined as the method of gathering knowledge from the process of answering one question, and subsequently using that knowledge to solve another adjacent problem (Tida and Hsu, 2022). Transfer Learning helps us leverage work done by large organizations. This approach provides the benefit of having a reliable output that comes from training on a much larger dataset and transferring that information to a problem that would otherwise have a much more limited amount of data. Transfer Learning allows us to bypass the issue of limited data by introducing additional information into the adjacent problem, and ultimately leading to improved results.

As the analysis done on fake news and fake reviews tends to differ quite significantly, it may be interesting to explore the impacts of some of the methods employed in the detection of fake news. While some work has been done whereby word embeddings have been trained on the data and then used as an embedding layer, there appears to be little work done in applying pretrained word embeddings to the same problem. In the following sections, I have explored the application of pretrained word embeddings in the prediction of fake reviews.

3. PROPOSED APPROACH

Considering all the research done in this area, I am proposing the application of pretrained word-embeddings to the detection of deceptive reviews. Word embeddings are a tool for representing text such that words with similar meaning have a similar representation. As seen in figure 1 below, words that are related are closer together in the vector space.

Figure 1: Visualization of Word Embeddings (Winastwan 2020)

Word embeddings have been used to solve several NLP problems due to the capability of embeddings to provide additional information about the “semantic properties and linguistic relationships” between words in the text (Wang et al., 2018). Word embeddings have been “commonly leveraged as feature inputs to downstream machine learning models” (Wang et al., 2018). These vector representations have led to the development of more robust models; however, there has been little work on using word embeddings in fake review detection.

“Traditional (count-based) feature engineering strategies” for text have been effective for extracting features from text (Sarkar, 2019). Most of these models are a bag of unstructured words and are missing information about the meanings of the words, the context of nearby words, as well as the structure and order of the text. Word embeddings are an enhancement on the traditional models as they can capture information on semantics and structure. Vector representations thus provide an improvement to the features that will be used to train the models.

The proposed approach is to use publicly available word embeddings that have been trained on very large datasets. The computationally expensive training has been completed in the external development of these word embeddings, and my analysis uses the knowledge captured in the embeddings to introduce some additional information to customer reviews. By applying word embeddings, the features will include information regarding the semantic meaning of the words. GloVe, Word2Vec and BERT are the three word embeddings that were applied in my analysis.

4. METHODOLOGY

This research paper applies machine learning and natural language processing methods to the textual component of a review. While the inclusion of numerical features such as rating may be useful to generate a more robust model, I am focusing my analysis on the text component for the purpose of understanding what valuable information can be obtained from language.

As shown in figure 3 below, a systematic approach was taken to understand how the tokenization of text can improve the predictions obtained by various classifiers. First and foremost, I began by extensively reviewing and cleaning the data. At this preliminary stage, I ensured that there was no missing or duplicated data and cleaned the text by doing several transformations on the words. To better understand the texts, I performed an extensive exploration of the data to understand how each feature relates to the target variable.

The text was then tokenized so that the words could be used as features on which to train the classification models. I applied a few traditional models to provide a baseline against which to compare the performance once word embeddings were applied.

Figure 2 below shows the visual representation of my methodological approach. The red box shows the process of developing a pretrained word embedding from a much larger corpus. The development of these word embeddings is outside the scope of this paper. The blue box shows my analysis whereby I apply the knowledge obtained from the pretrained word embeddings to change the feature space in such a way that it contains the semantic meaning of the words.

Finally, the data is clean and ready to have the machine learning models trained and tested. In my analysis, I have applied 8 machine learning models and assessed their results.

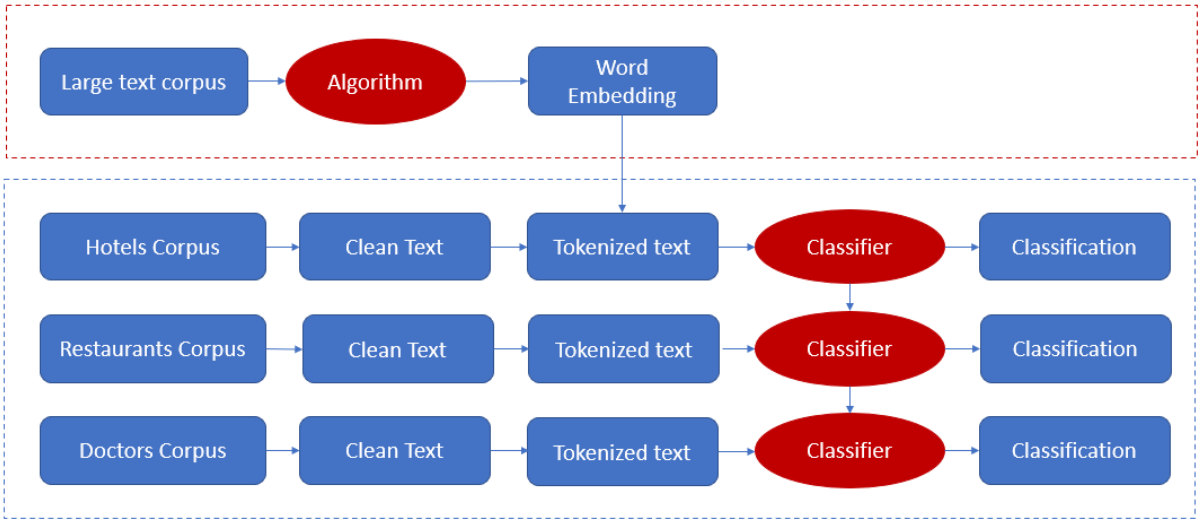

Figure 2: Visual representation of the methodology applied in this paper

4.1. DATA UNDERSTANDING/CORPUS

The following analysis was completed on online customer reviews of services, specifically opinion spam in hotel, restaurant, and doctor reviews. The Deceptive opinion spam dataset is a corpus that consists of truthful and deceptive hotel reviews of 20 Chicago hotels. The corpus contains: 800 truthful reviews and 800 deceptive reviews including domain expert deceptive opinion spam from employees, crowdsourced deceptive opinion spam from Mechanical Turk, as well as truthful Customer reviews. (Li et al., 2014)

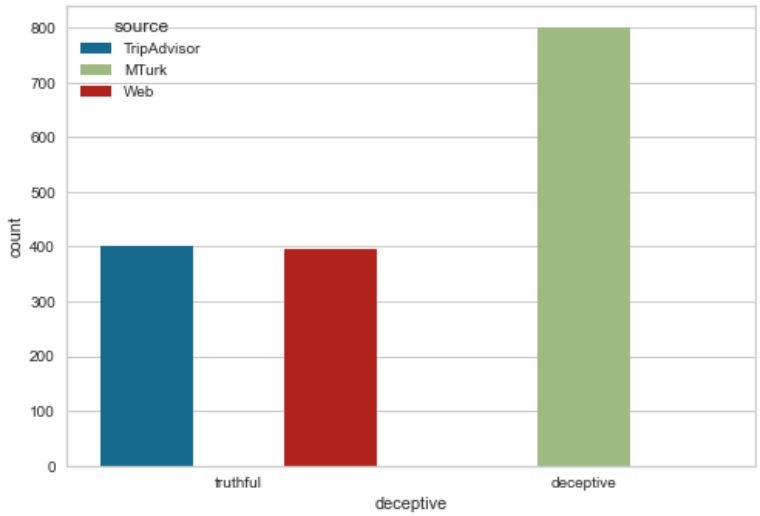

Figure 3: Breakdown of truthful and deceptive reviews across sources in the corpus

A cross domain gold standard dataset was created for Restaurants and Doctors with 400 and 558 reviews respectively. These data sets contain predominantly positive reviews. The expectation is that the performance of the model on hotel reviews and restaurant reviews is quite similar due to the many shared language between domains. Alternatively, the performance on the doctor dataset should suffer due to differing vocabulary. (Li et al., 2014)

4.2. DATA PREPARATION

To clean the data, the following processing steps were done. I fixed spelling errors using TextBlob and removed 4 duplicates in the hotel dataset. HTML tags, numerical data and punctuation were removed. Text was lowered and lemmatized, and I removed stop words and words with a length less than 3. Regarding the deletion of stop-words, I used the stop words provided by the NLTK library for English.

The most crucial step needed to be able to feed the data to machine learning models is the tokenization of the text. Computers cannot understand text so this step is required to transform text into a vector of numbers that can be used as features in the models. Every word is a feature with a numerical value. As a result, the feature space is defined by the vocabulary.

This paper aims to compare various data preprocessing steps. The data cleaning step will be kept consistent so that the analysis can focus on the tokenization step. The goal is to understand how much of an impact different tokenization steps can have on the overall performance of the machine learning models applied. I have reviewed some traditional tokenization steps in the following section.

4.2.1. Bag of Words

The first tokenization method reviewed was bag-of-words, which was treated as our baseline model to compare our results against. Bag of words is a tool that shows which words occurred within a document. It develops a vocabulary that includes all the words present in a text, as well as a measure of the occurrence of those words (Brownlee, 2017).

It is an unstructured group of words as the model does not account for the order or meaning of the vocabulary within the document. The model recognizes documents that have the same content as being similar. It is important to note that the model will not recognize two synonyms as being analogous as the model does not include any information regarding the semantic meaning. The vector representation has a length equal to the size of the vocabulary. As each document only contains a small subset of the vocabulary, the resulting vector is very sparse. (Brownlee, 2017)

Figure 4: Example of a bag of words for one document

There are some challenges that arise when using Bag-of-words. Firstly, it has the curse of dimensionality issue as the total dimension is the vocabulary size. As such, it can easily over-fit your model. Secondly, processing sparse vectors can be very computationally intensive as they require a lot of memory. Lastly, Bag of words representation doesn't consider the relationship between words. As mentioned above, there is no record for which words appear together. (Brownlee, 2017)

4.2.2. TF-IDF

TF-IDF is a composite score that accounts for how unique or common a word is. The model rescales the frequency of words by how often they appear in all documents, reflecting the ability of a word to uniquely identify the document (Brownlee, 2017). "It represents the composite weight of each term in a document." (Antonio et al., 2018).

Term Frequency (TF) is a scoring of the frequency of the word in the current document and Inverse Document Frequency (IDF) is a scoring of how rare the word is across documents (Antonio et al., 2018). As shown in the computation below, the resulting composite score is calculated by multiplying TF and IDF. If a word is very common then IDF is near to zero, otherwise, it is close to 1. The higher the TF-IDF value of a word, the more unique/rare occurring that word is. If the TF-IDF is close to zero, it means the word is very commonly used. (Qaiser and Ali, 2018)

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

The scores effectively highlight the words that contain useful information in each document.

4.2.3. POS-tagging

	clean_text
0	[(stayed, VBN), (night, NN), (gateway, NN), (f...
1	[(triple, JJ), (rate, NN), (upgrade, JJ), (vie...
2	[(come, JJ), (little, JJ), (late, JJ), (finall...
3	[(igni, NN), (chicago, NN), (delivers, NNS), (...
4	[(asked, VBN), (high, JJ), (floor, NN), (away,...

Table 2: Example of POS-tagging for 5 records of our corpus

Part-Of-Speech Tagging identifies the function of each word or character in a sentence or paragraph. Words in a text are categorized in correspondence with a particular part of speech, depending on the definition of the word and its context. The words are then given a "tag" according to the category they fall into. This additional information is used by the model to improve classifications. (Kumawat and Jain, 2015).

BoW, TF-IDF and POS-tagging were the traditional models reviewed in this paper. In the following section, I will explain the advanced language models that were assessed in my analysis: GloVe, Word2Vec and BERT word embeddings.

4.2.4. GloVe

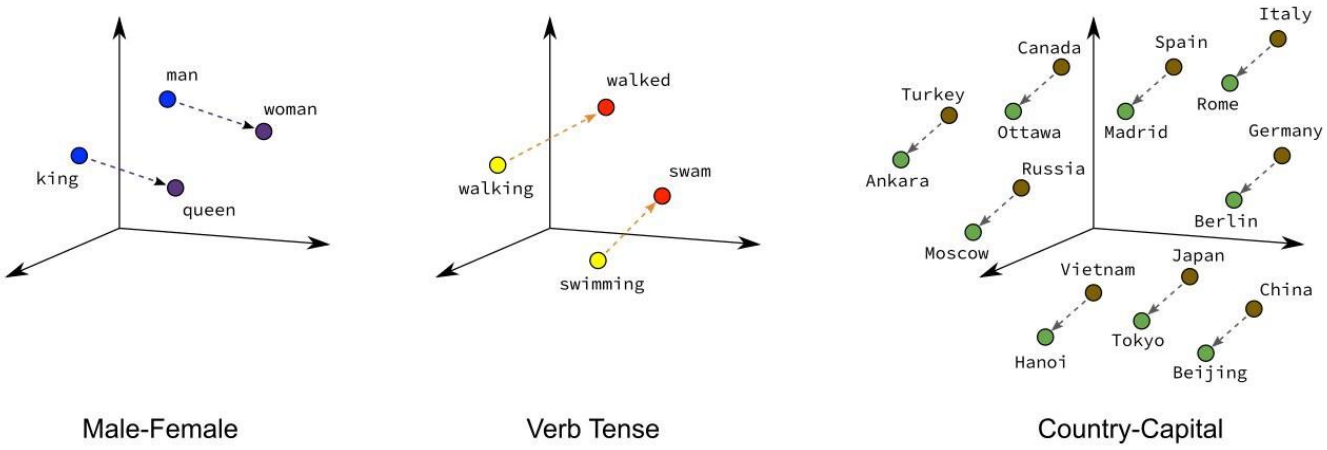

Figure 5: Visual representation of word embeddings (Khandelwal, 2019)

GloVe is an unsupervised learning algorithm that was trained on a corpus of 6 billion words from Wikipedia 2014 and Gigaword 5. "Training is performed on aggregated global word-word co-occurrence statistics from a corpus, and the resulting representations showcase interesting linear substructures of the word vector space" (Pennington et al., 2014)

The GloVe model tabulates how frequently words occur together in each document. It is “a log-bilinear model with a weighted least-squares objective” (Pennington et al., 2014). The model essentially obtains information about the semantic meaning of the words based on the “ratios of word-word co-occurrence probabilities”. The resulting embedding reflects the probabilities that two words appear together. (Pennington et al., 2014).

The idea of nearest neighbours is used to map words into a vector space in such a way that the Euclidean distance (or cosine similarity) between two vectors measures the linguistic or semantic similarity of the words. That is, the closer the word vectors are to one another, the more similar is their meaning. (Pennington et al., 2014).

4.2.5. Word2Vec

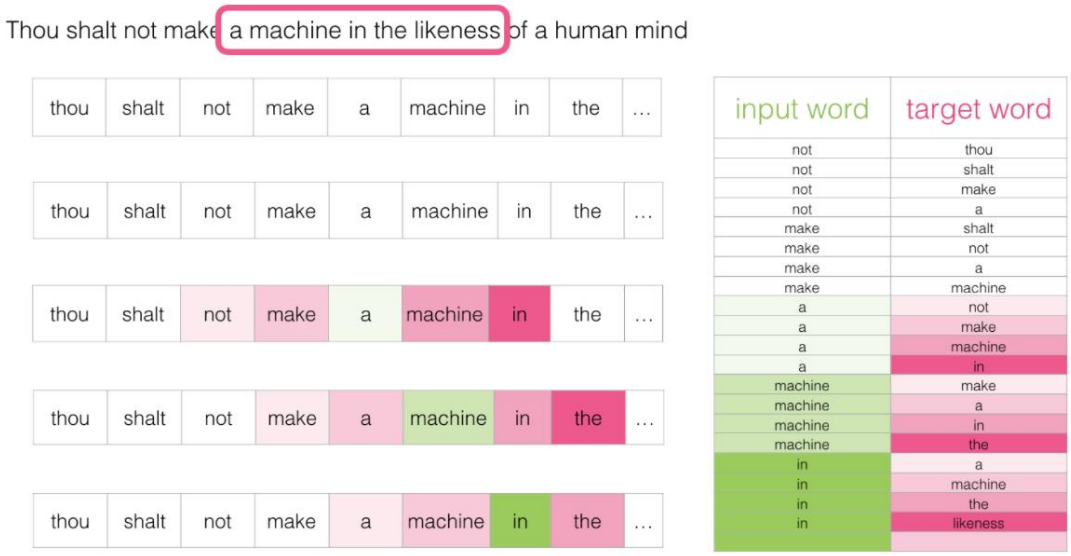

Figure 6: Visual representation of Word2Vec’s skip-gram model (Alammar, 2019)

Word2Vec is a method of developing pretrained word embeddings by training on a neural network to learn relationships between words. There are two Word2Vec models that have been produced: Continuous Bag-of-Words (CBOW) and Skip-gram. CBOW is a model that has been trained to be able to predict the missing word in a sentence. The model guesses a missing word based on the words before and after it (i.e., the context). (Alammar, 2019)

In contrast, the Skip-gram model learns which words are typically near each other and can predict which words are likely to be near a certain input word. Given a specific word in the middle of a sentence (the input word), the model looks at the words nearby. The resulting output is the probability of every word in our vocabulary of being the “nearby word”. In other words, how likely it is to find each vocabulary word near our input word. (Alammar, 2019). We have applied Word2Vec’s Skip-gram model in our analysis.

4.2.6. BERT

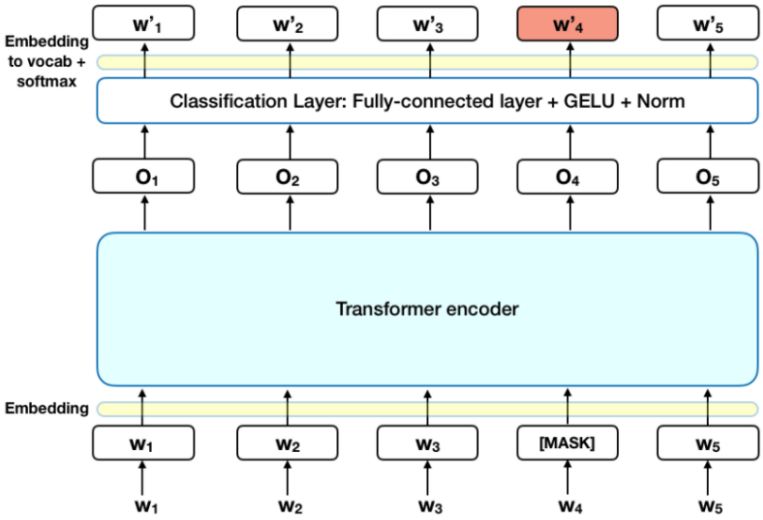

Figure 7: Visual representation of BERT algorithm (Liang, 2022)

The last word embedding that we assessed was Bidirectional Encoder Representations from Transformers (BERT). This model implements an attention mechanism that learns contextual relations between words in a text, also known as a Transformer (Devlin et al., 2018). “The Transformer includes two separate mechanisms — an encoder that reads the text input and a decoder that produces a prediction for the task”. Only the encoder mechanism is necessary in our case (i.e., to generate a language model) (Devlin et al., 2018).

BERT produces word representations that are dynamically informed by the surrounding words. That is, the model takes into consideration the context within which a word appears. (Devlin et al., 2018). This is an advantage over GloVe and Word2Vec where each word has a constant representation, regardless of the surrounding words. We expect BERT to outperform the other models due to the more accurate representation of words.

All the vectors produced by each word embedding will be used as feature inputs to downstream machine learning models. I have covered the downstream application in the following section.

4.3. MODELING

Once our data was cleaned and preprocessed, I began to apply the different types of models, with varying parameters, to see how the performance on the validation set changed. With each model I set a baseline with default parameters. The models tested are as follows:

Simple Logistic Regression (LR) for classification was applied. This model aims to minimize the sum of squared errors between the modelled values and the real values. GridSearch was applied to obtain the parameters that optimize the performance of the model. Once I obtained the optimal set of parameters, I initialized a model with those specifications and fit the training data on it.

Stochastic Gradient Descent (SGDC) is an iterative model for optimizing the objective function. As it estimates the gradient from a random subset of the dataset, "it can be regarded as a stochastic approximation of gradient descent optimization" (Bottou, Bousquet, 2012). SGDG is often applied to sparse or highly dimensional machine learning problems which is typical of text classification and natural language processing.

K nearest neighbors (KNN) is a non-parametric model which uses proximity between data points to obtain a probability that a data point behaves like its neighbors (Soucy and Mineau, 2012). GridSearch was used to obtain the optimal number of neighbors, the algorithm to compute the nearest neighbors, as well as the appropriate weighting of those neighbors.

Support Vector Machine (SVM) is a non-probabilistic binary linear classifier that was also considered (Cortes and Vapnik, 1995). The algorithm maps records into a space such that the width between the two categories is maximized. One parameter I reviewed was C, which is a regularization factor and states how much misclassification you are willing to accept. I also looked at the impact of different kernels, along with their corresponding gamma.

The following Decision Tree models were applied: Simple decision tree (DT), Random Forest (RF), Gradient boosted tree (GBT) and XGBoost (XGB) (Song and Lu, 2015; Breiman, 2001; Bentejec et al., 2015). These models learn simple decision rules from the training data. As Decision Trees are known to overfit, we have reviewed RF, GBT and XGB which use various generalization techniques. Random Forest is an ensemble of Decision Trees, while XGBoost uses advanced regularization which improves model generalization capabilities.

To start, I identified a baseline model for each classification model, with all parameters set to default. Different runs were done on each model whereby one parameter was changed, while holding the others constant (ceteris paribus). Considering the performance change on the validation set I got an idea of impactful parameters and their possible range of values. Having gained an overview of the important parameters and their ranges, I deployed a grid search algorithm to find the optimal set of parameters. Having identified the best possible parameters, I trained a final model based on those parameters.

4.4. EVALUATION

To compare the methods assessed, an evaluation method and success criteria need to be defined. The selection of machine learning models was trained, and their performance was evaluated using cross-validation (CV). CV allows for the development of generalizable models

that are not over fit to one dataset. K-fold CV works by randomly partitioning data into k samples (Bates et al., 2021). In this study, the dataset was divided into 10 folds. Each of the 10 folds was used as a test set and the remaining nine were used as training data. Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) were calculated for each of the 10 folds, and the global mean (μ) and standard deviation (s) were used to assess each method's performance. (Bates et al., 2021)

While evaluating the experimental results, I considered several evaluation measures. Accuracy was used to provide a broad evaluation of performance. It represents the percentage of reviews that are correctly classified (Vujovic, 2021). In our case, deceptive reviews have been labeled as 1, while truthful reviews have the label 0. As such, the False Positive (FP) rate stands for the percentage of truthful reviews incorrectly classified as deceptive, while False Negative (FN) rate is the percentage of deceptive reviews misclassified as truthful. As deceptive reviews are harmful, it is essential to correctly classify those reviews. That is, it is important to obtain a lower false negative rate. We would be aiming to capture all the deceptive reviews, even if this means that some truthful reviews are misclassified.

It is also interesting to look at Precision and Recall. Precision is the proportion of deceptive classifications that were correct, and Recall is the proportion of deceptive reviews that were identified correctly. As mentioned above, we will focus our evaluation on identifying as many deceptive reviews as possible. In this case, obtaining a high Recall is more important than high Precision.

In addition to the numerical evaluations, I also applied a couple of data visualization methods. Being able to view all the data at once enhances data understanding. As text data notoriously has many dimensions due to the nature of language, it is quite challenging to visualize textual data. To address these limitations, I used a couple tools that were built purely for this purpose.

Uniform Manifold Approximation and Projection (UMAP) is a tool for visualizing highly dimensional data by optimizing a low-dimensional graph such that it effectively preserves the global structure of the data. UMAP is a very effective visualization tool as it has no computational restrictions on embedding dimensions. (McInnes and Healy, 2018).

Finally, I used the SHapley Additive exPlanations (SHAP) library which uses game theory to explain how each feature affects the model. "SHAP values are used to increase transparency and interpretability of machine learning models" (Lundberg and Lee, 2017). It shows the contribution of each feature on the final predictions of the model; however, it does not evaluate the quality of the prediction. The basis of this library is the Shapley Value which is the average expected marginal contribution of one player (i.e., feature) after all possible combinations have been considered. (Lundberg and Lee, 2017)

5. ANALYSIS OF EXISTING TEXTS

5.1. EXPLORATORY DATA ANALYSIS

I explored some features of the dataset to see if I could find any glaring differences. In figure 8, I have created two plots showing the word count for deceptive and truthful reviews before and after the text was processed. It appears that truthful reviews have a higher peak showing that they tend to be shorter. Otherwise, the patterns between deceptive and truthful reviews appear to be quite similar.

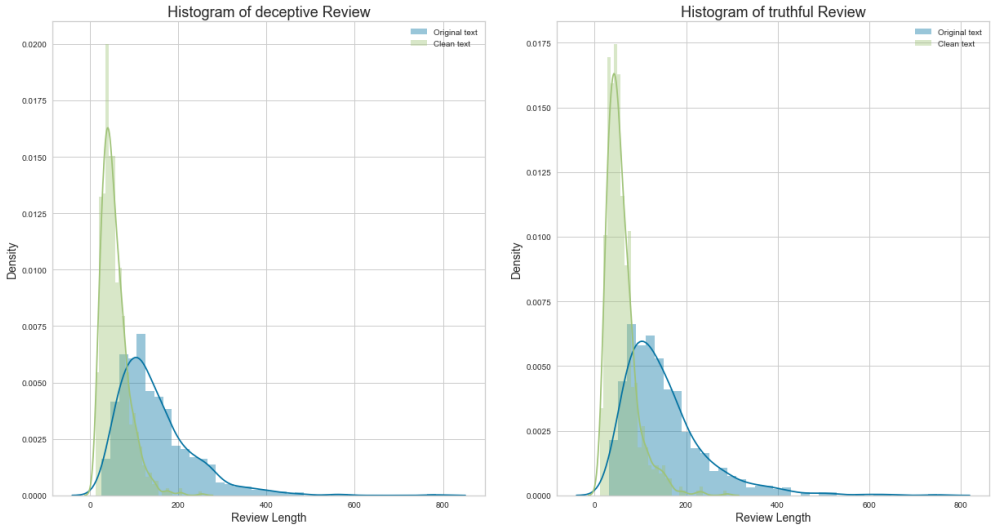

Figure 8: Histogram showing the word count of original and clean text by label

I completed a preliminary analysis of the text by computing the total word frequency for each set of texts. Figure 9 shows a matrix of word clouds that display words that are most often found within each category of reviews. The larger the words, the more frequent they are in those reviews. It is easy to identify differences in the language used in truthful and deceptive reviews. The positive reviews seem to be quite similar with a high focus on the hotel staff. The truthful ones have a slightly higher focus on the location. Negative reviews seem to differ somewhat, although there appears to be a higher focus on the words “night” and “desk”.

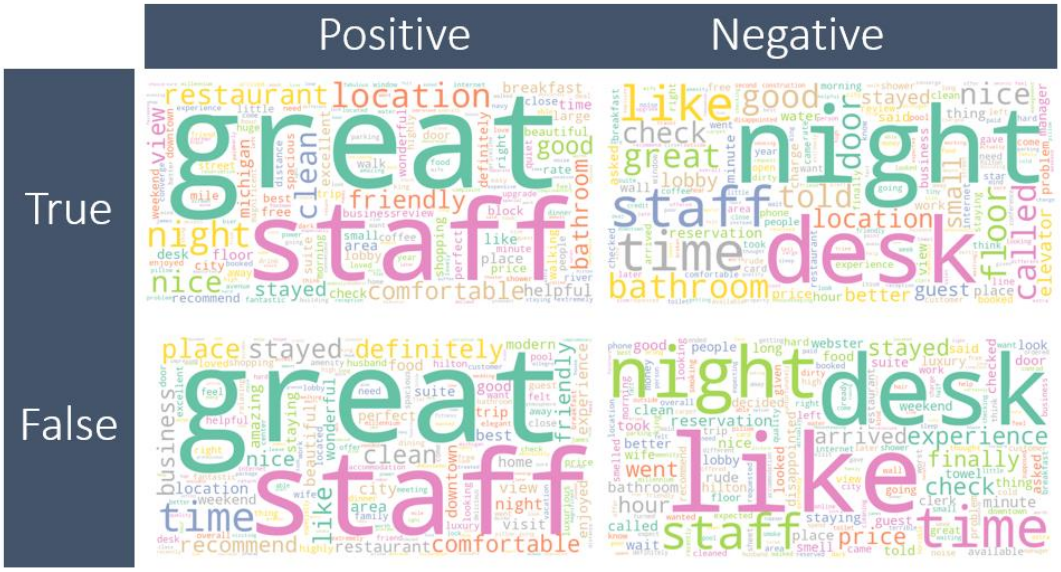

Figure 9: Word clouds by label and polarity

5.2. SENTIMENT ANALYSIS

Sentiment analysis is the detection of a consumer's level of satisfaction based on their review. In cases where a numerical rating is not made, it can be difficult to identify a consumer's sentiment. Sentiment Analysis is the automatic detection of the polarity of a review based on the type of words used in the review (Alaei et al., 2019). Sentiment analysis is a form or transfer learning as it uses a pretrained model that has learned the sentiment of certain words and is able to apply the model to detect the sentiment of a new text. I have used the Hugging Face library to develop the Sentiment analysis for the hotels, restaurant and doctor data presented in Figure 10 below.

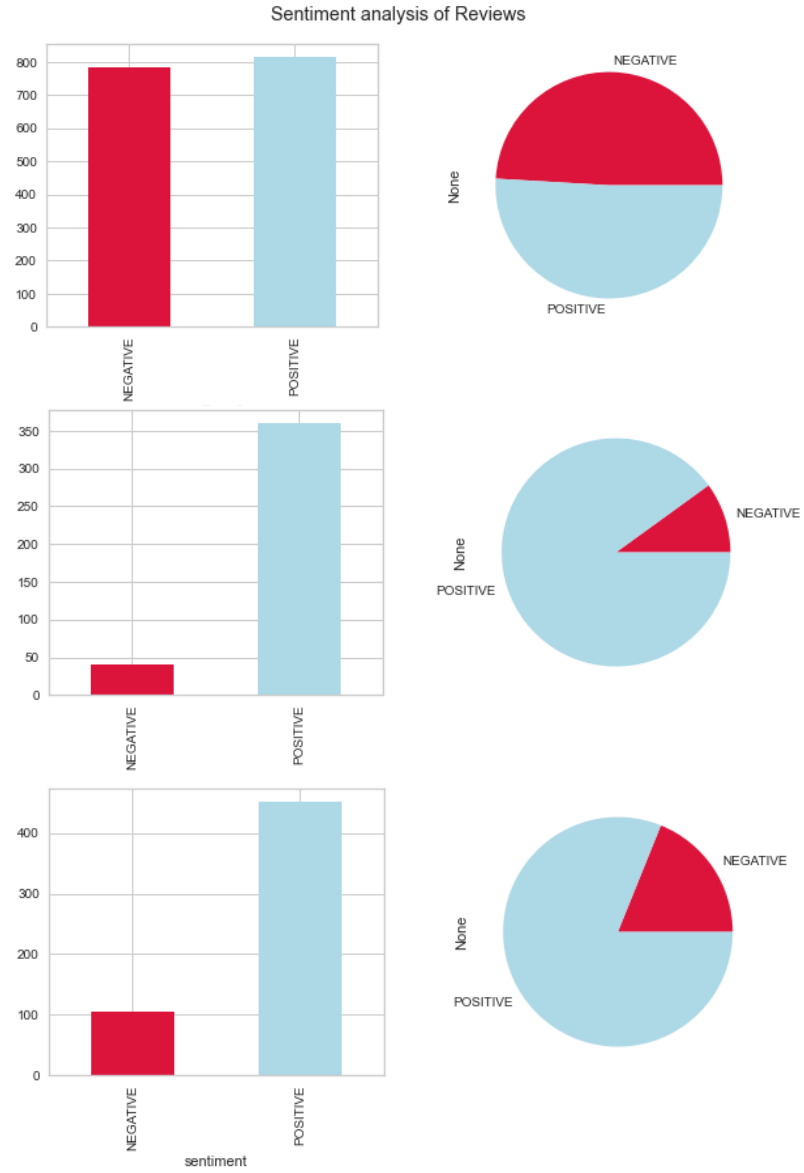

Figure 10: Sentiment Analysis of Hotel, Restaurant and Doctor reviews

Fake review detection research has mainly relied on textual and behavioral features. Metadata about the review may also be useful. Additional natural language features such as "semantic similarity and emotion, a wide variety of lexical and syntactic features and deeper details such as understandability, level of details, writing style and cognition indicators" have been proposed. (Barbado et al., 2019)

5.3. TOPIC MODELLING

Topic Modelling is a statistical technique for building clusters of words. The texts within the dataset are thereby “a mixture of all the topics, with each topic having a different weight”. This results in the main themes being extracted from the dataset (Blei et al., 2003). I thought it would be interesting to explore the topics found in the dataset to see if there are any obvious differences.

Topic Models can uncover the hidden structure of a group of texts. Latent Dirichlet Allocation (LDA) is one such model and was applied to our corpus. The “generative probabilistic model assumes each topic is a mixture over an underlying set of words, and each document is a mixture of over a set of topic probabilities.” (Blei et al., 2003). The model essentially builds clusters of words, and each text is then assigned each word cluster with a certain probability.

As can be seen in the figure below, the distances between the various topics are minimal with only 3 topics that have a higher distance from the average. I also noticed that the top relevant words seem to be quite consistent across topics. As can be seen, the words room, hotel and Chicago can be found in most reviews.

With the short text and the generally similar text, the features do not vary significantly between reviews. Perhaps some traditional text mining methods may be able to capture differences that cannot be seen by the naked eye.

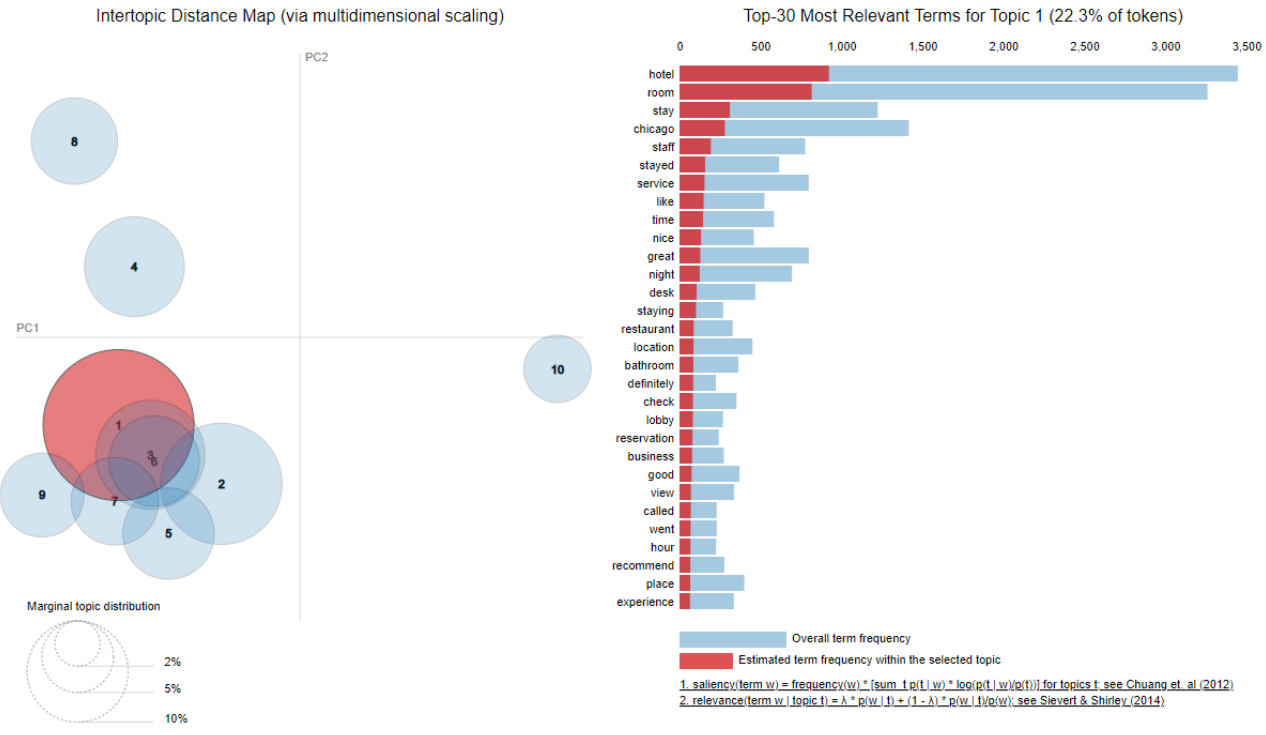

Figure 11: LDA topic analysis

The marginal topic distribution refers to the percentage that the topic makes up in the corpus. This figure shows how topics relate in a 2-dimensional space based on principal components PC1 and PC2.

5.4. BAG OF WORDS

I applied Bag of Words to review the top 20 unigrams, bigrams, and trigrams. The figure below shows the top 20 bigrams in the corpus. As you can see, room service is the most common bigram, meaning a lot of reviews may be commenting on this. Things like hotel room and customer service are bigrams that we would expect to see.

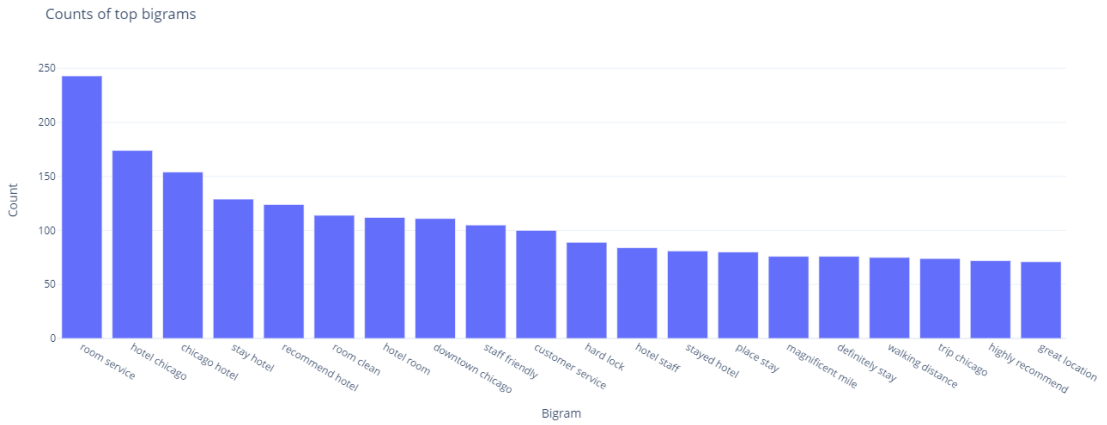

Figure 12: Counts of top bigrams found in the corpus

In our preliminary analysis, I found that the inclusion of higher order n-grams led to artificially low scores. This makes sense as the texts within the corpus are not particularly long, and most of the models that have been tested rely on the count of times these n-grams are present in the corpus. As such, I only considered 1-grams in every model applied.

6. RESULTS AND DISCUSSION

6.1. TRADITIONAL MODELS

In figures 13 and 14 below, you can see the results obtained on the BoW model. The performance is quite good on most models with very tight box plots between 0.8 and 0.9. The best performing model here is logistic regression with an average value of 86.28%. The model was also applied to the doctor and restaurant datasets. Apart from KNN, the model performed the best on the hotel's dataset for every model. Interestingly, the model had superior performance on the doctor's reviews, over the restaurant reviews. KNN and SGDC had higher performance on the restaurant dataset, compared to the doctor's dataset. Except for KNN and DT, all the models obtained an average Precision above 0.9.

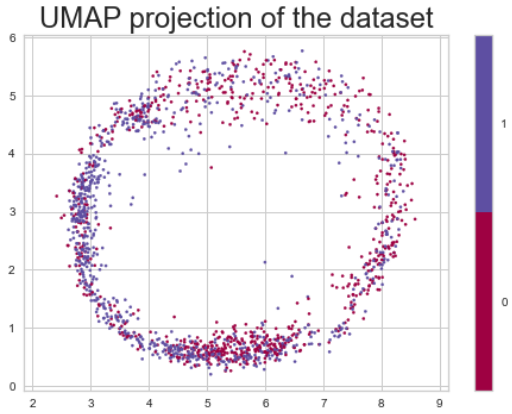

Figure 13: UMAP projection of BOW model

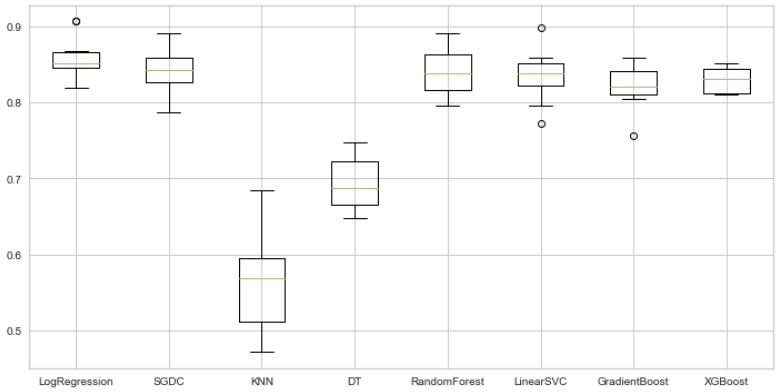

Figure 14: Box plots of the performance of all classifiers on the BOWs model

The TF-IDF results follow a similar pattern to the BoW results that were just reviewed. Logistic Regression once again with the highest result at 84.87%, followed closely by RF and SVM at 84.33% and 83.62%. All models except for KNN are minimally trailing the BoW results and the box plots are somewhat larger than the ones we are seeing for BoW. Similarly, the average Precision for most models is also around 0.9. The UMAP projection is quite interesting as there seems to be a clearer separation of the true and fake reviews.

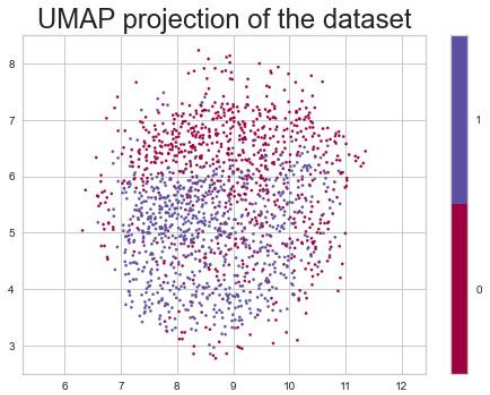

Figure 15: UMAP projection of TF-IDF model

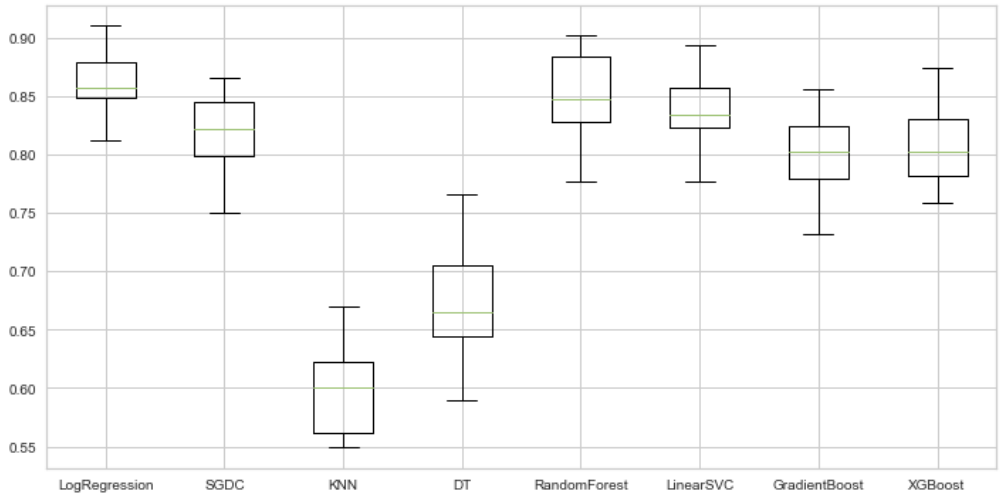

Figure 16: Box plots of the performance of all classifiers on the BOWs model

The POS-tagging results also show tight box plots with the same models performing well. SGDC, KNN and SVM appear to be performing somewhat better in this case. Generally, KNN and DT are significantly trailing the other models. The poor performance of KNN seems to make sense as all reviews are very similar to one another with no obvious distinctions.

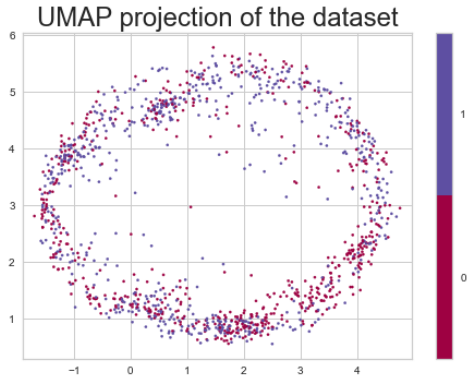

Figure 17: UMAP projection of POS-tagging model

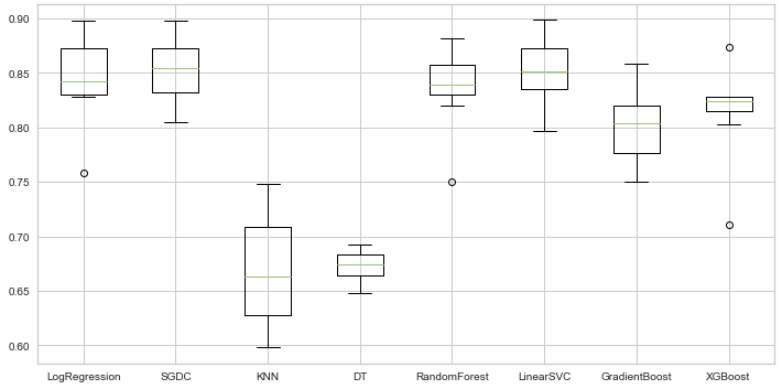

Figure 18: Box plots of the performance of all classifiers on the POS-tagging model

Next let's look at the results obtained by applying word embeddings in the preprocessing of the text.

6.2. WORD EMBEDDINGS

Below you will see the results obtained by using GloVe pretrained word embeddings. Interestingly the results seem to trail the traditional models reviewed above, and the box plots show some higher variability. As can be seen in the UMAP, it is very difficult to differentiate between the dots. As Linear Regression had the best performance, I tuned the parameters using GridSearch and was able to obtain a score of ~80%. While this is a 10% improvement to the score, it still trails the results obtained by the traditional language models.

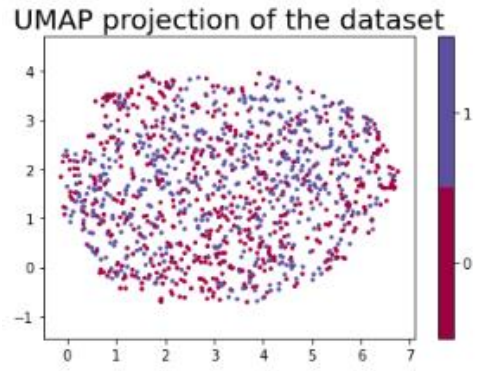

Figure 19: UMAP projection of GloVe model

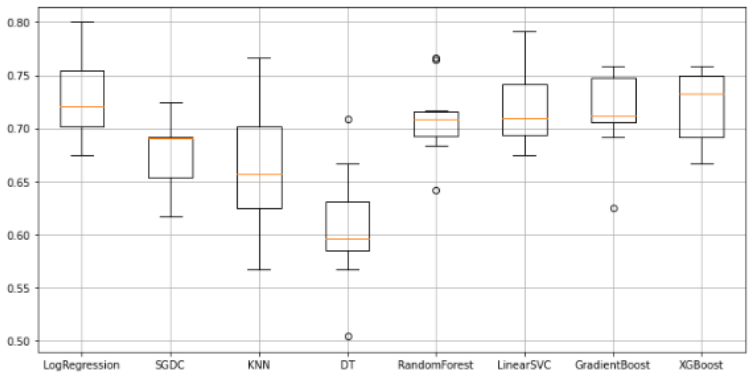

Figure 20: Box plots of the performance of all classifiers on the GloVe model

The results obtained by Word2Vec skip gram seem to show a very slight improvement over GloVe but the results are still lagging our traditional models. Logistic Regression is the front runner yet again, although the box plot has a much larger range. Like GloVe, the UMAP does not show distinct differences.

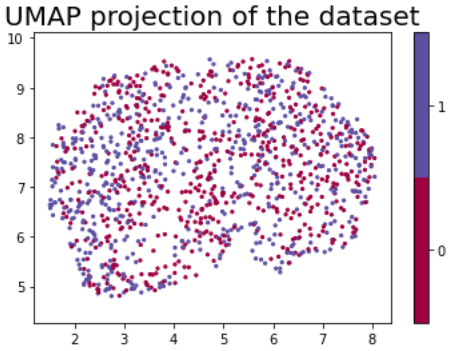

Figure 21: UMAP projection of Word2Vec model

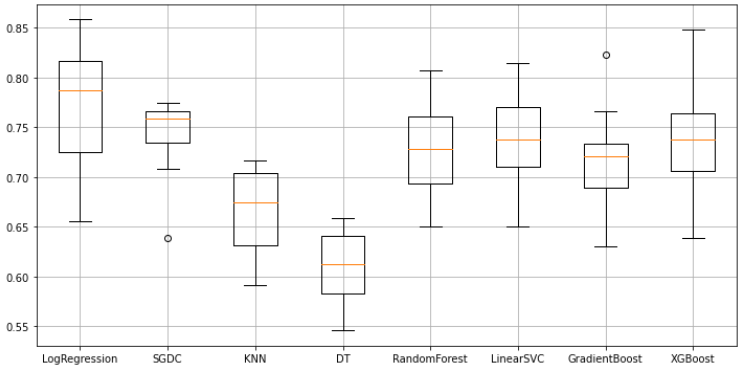

Figure 22: Box plots of the performance of all classifiers on the Word2Vec model

Finally, the box plots are a bit tighter for BERT meaning the results tend to be more consistent; however, the performance is still lower than other models reviewed. One thing to note is that the variability across all the models is lower when the word embeddings have been applied (i.e., the range between the highest and lowest scores achieved across all models is lower). That said, we are seeing decreased performance in the highest score obtained by the word embeddings.

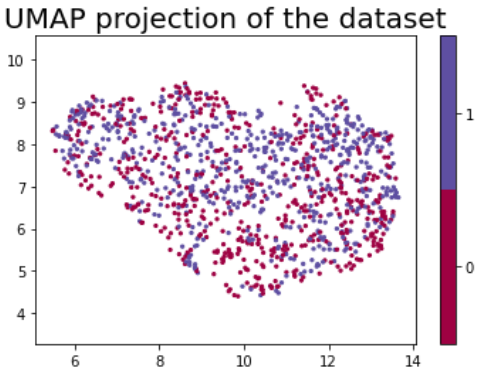

Figure 23: UMAP projection of BERT model

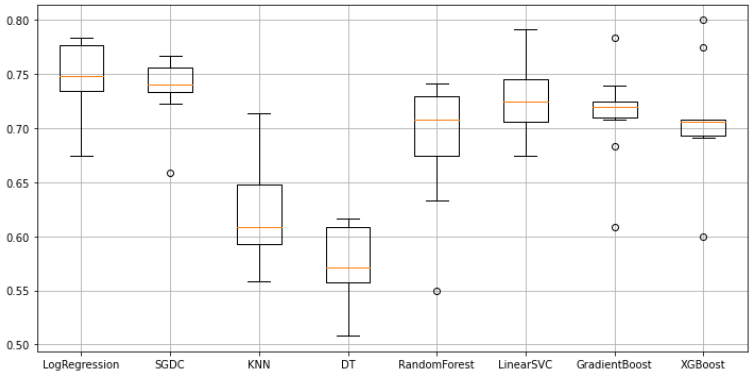

Figure 24: Box plots of the performance of all classifiers on the BERT model

Perhaps the decreased performance can be explained by the fact that adding semantics and context to the text leads to losing the relationships between the words that are found in the text, which may have more predictive ability than the context of a word.

To test this hypothesis, the word embeddings were applied to the restaurant and doctor datasets, to see if they were able obtain improved results. In figures 25, 26 and 27 below, you can see the performance of each word embedding and model pair for the hotel, restaurant and doctor data sets.

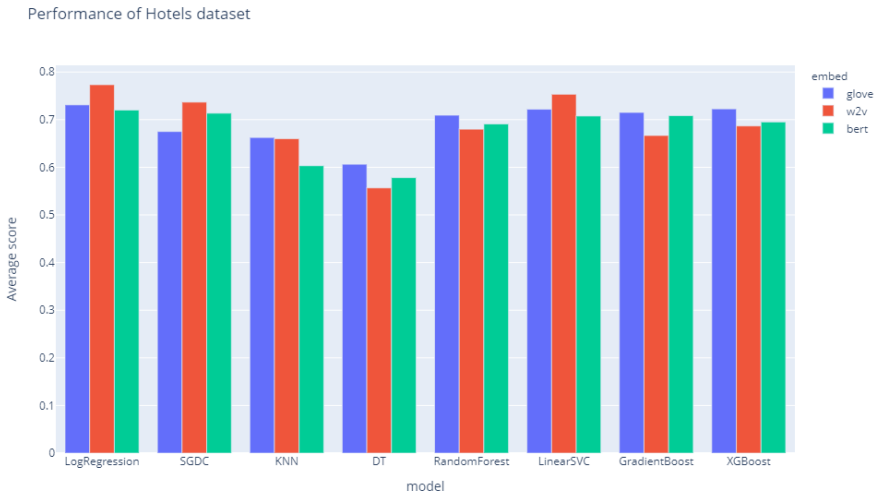

Figure 25: Performance of all models for each word embedding applied on the hotels dataset

Interestingly, the performance of GloVe and Word2Vec appears to be higher on the doctor dataset, across all models trained. We would expect to see similar performance of the models for hotels and restaurants, as the content language is more similar. Perhaps there is more variability in the language used in the doctor reviews, leading to clearer patterns for the algorithms to identify. It appears that BERT had quite similar performance across all datasets and models. It seems that BERT may be a superior pretrained word embedding as the results are more reliable and consistent. Although BERT had lower performance on the hotel review dataset.

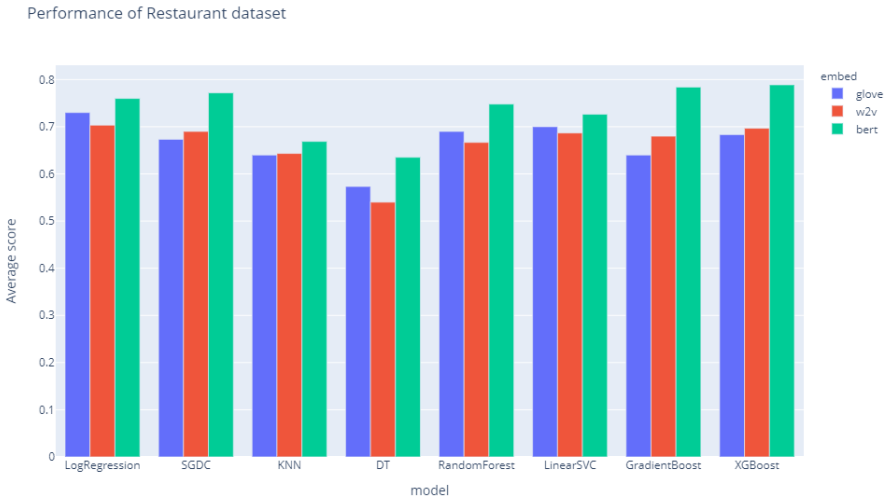

Figure 26: Performance of all models for each word embedding applied on the restaurant dataset

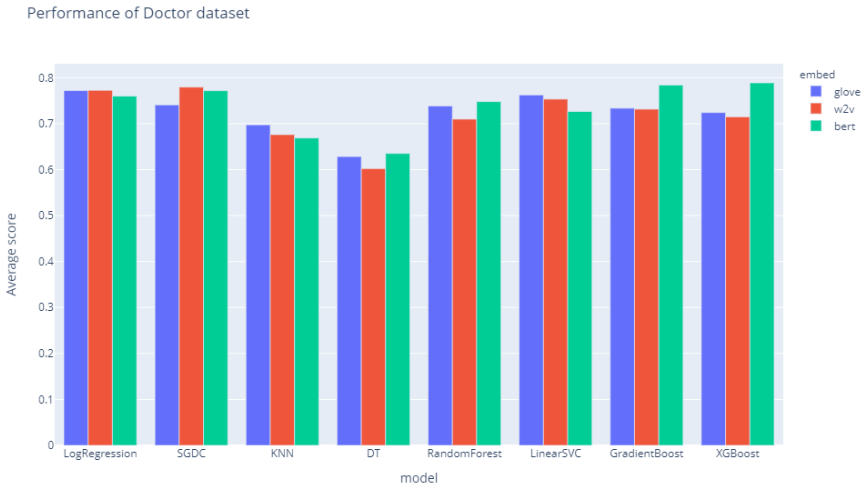

Figure 27: Performance of all models for each word embedding applied on the doctor dataset

Now that I have assessed the various models, the model that obtained the best performance was our baseline Bag-of-Words with simple Linear Regression. As the results of machine learning models can often be challenging to interpret, I have used SHAP to help explain the patterns learned by these models. In this way, one can understand how the features affected the performance of the models. Below you can see how the selected model should be interpreted.

In the following figure on the left, you will see the SHAP values for a single review. The x-axis has the values of the dependent variable (i.e., whether the review is deceptive or not). F(x) is the value predicted by the model. The SHAP value for each feature is given by the length of the bar. The absolute SHAP value shows how each feature affected the prediction. Please note that in our case a lower value is representative of a truthful review and a higher value of a deceptive one. For the review shown, “hotel”, “helpful” and “rate” were indicative of a truthful review while “money” and “clean” would appear to have a negative impact on the prediction made. These seem to be reasonable patterns for the model to have learned. These SHAP values for each word are valid only for this observation. The same word may have a different

SHAP value in another review. Similarly, the plot on the right shows the main features affecting the prediction of a single observation.

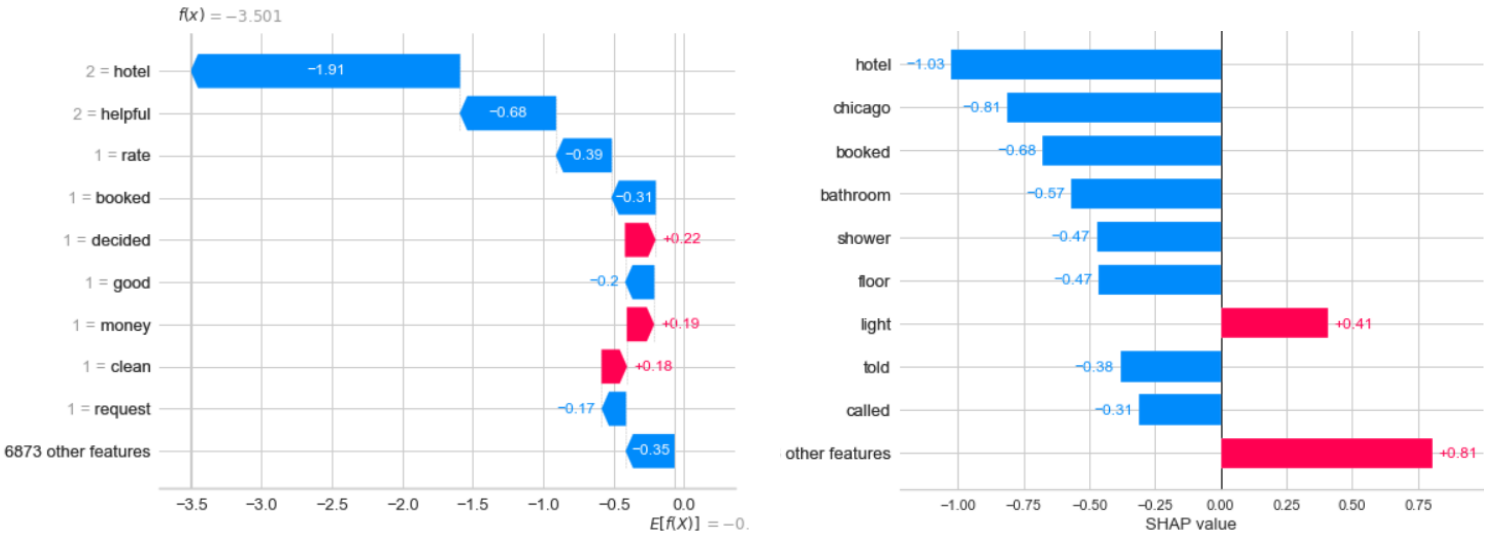

Figure 28: Visualization of the SHAP values for one observation

The plot below shows the global effect of the features. Here, the features are ordered from highest to lowest effect on the prediction in absolute terms. Each dot represents a single observation. Since our models have a large feature space, it is difficult to directly visualize and the interpretability is affected. These visualizations are incredibly useful to understand the patterns learned by black box models.

Figure 29: Visualization of the global effect of the features on the predicted classification

7. CONCLUSION

Given the short length of reviews, all the models applied performed reasonably well. There is some interesting information that we have learned from the small differences that can be seen across pre-processing steps and models applied.

As I noted in the exploratory data analysis, the hotel reviews contain very similar words. We can understand why the bag of words and TF-IDF models would perform well as they consider the occurrence and frequency of words across documents. POS tagging also accounts for how the words are used and provides an additional element of information to the models. It is interesting to note how well the traditional models can differentiate between deceptive and truthful reviews.

Word embeddings add an element of the meaning of the words and provide more context for the algorithms. However, with the introduction of word embeddings, we lose the relationship between the words that are present in the text. Given that the word embeddings had relatively weaker performance, I can conclude that the presence of certain words is a better predictor of deceptive reviews than the semantic meaning of the words. This makes sense as there is not a lot of variability in review texts. Reviews that are reflecting a similar sentiment may use different words that reflect the same meaning. Once the word embeddings are applied, these reviews would be very close to one another in the vector space since the overall idea being portrayed is the same. It then makes sense that word embeddings did not greatly improve the performance of the models. BERT appears to be the most reliable of the pretrained word embeddings reviewed, with extremely consistent results across all the datasets. BERT had better transferability across different domains.

Overall, the analysis presented has led to a deeper understanding of deceptive reviews within the service industry. However, there are several areas to explore further and multiple ways we can improve our analysis to obtain predictive models with superior performance.

8. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

The main issue with fake review detection is the absence of accurately labeled data. The inability to identify the fraudulent reviews makes it difficult to create a program to automate this process. As such, the general population should be educated on the prevalence of such reviews and remain critical of everything that they read. Businesses should also introduce processes to verify a customer's review (I.e., validate that they were in fact a customer of the hotel). Perhaps in the process of validating a review, we could also develop a more robust data set that could be further reviewed and tested. Of course, having more data will also lead to better models and therefore, better predictive ability.

While there needs to be some room for error that comes with generalizing a predictive model, we also need to consider what amount of error is acceptable. As mentioned earlier, some deceptive reviews are more harmful than others. A review that praises a company that already has a high average rating is not very damaging, while a review that goes against the consensus is much more destructive. One negative review is much more impactful than several positive reviews. So, a deceptive positive review for a company that has generally negative reviews also will not be very impactful. Perhaps focusing the analysis on reviews with an extremely negative sentiment when compared to the average would also yield some interesting results.

While natural language processing focuses on extracting information from text, there may be an opportunity to provide some additional context to the data by including other metadata about the reviews and/or reviewers. Variables such as date of the experience, date of the review, as well as information about the reviewer such as age and location may be able to inform the detection of truthful or deceptive reviews. Having access to the IP address would also help to identify individuals that are repeat offenders. While it may be very difficult to capture all deceptive reviews, making the barriers to entry higher may provide some safety/security for business owners while also providing an enhanced data set for additional research.

In addition, the experience being reviewed is also an important consideration. In this paper, you will note that there was not much variability in the words used. As shown earlier in the topic modelling, 7 out of 10 topics had a high frequency of the words "hotel", "room" and "stay". Perhaps the analysis could be improved by using word embeddings developed from a dataset specific to the hospitality industry. As we saw, the performance of the word embeddings was somewhat higher on the doctor dataset. Pretrained word embeddings may not be the best option for short texts about specific domains. Word embeddings may be more effective when looking at longer texts (i.e., more data), such as fake news.

One other recommendation for future analysis would be to look at the impact of word embeddings in combination with Neural Networks – specifically Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM). These models consider the order of the words which may be a much stronger predictor than simply the existence of the word. Since many of the reviews in our corpus have very similar words, I expect that adding the element of order would produce higher performance.

9. REFERENCES

Alaei, A. R., Becken, S., & Stantic, B. (2019). Sentiment Analysis in Tourism: Capitalizing on Big Data. Journal of Travel Research, 58(2), 175–191. https://doi.org/10.1177/0047287517747753 IEEE Staff, & IEEE Staff. (n.d.). *2010 International Conference on Signal and Image Processing.*

Alammar, J. (2019). *The Illustrated Word2vec*. Jay Alammar. https://jalammar.github.io/illustrated-word2vec/

Antonio, N., Almeida, A. de and Nunes, L. (2016), "Using data science to predict hotel booking cancellations", in Vasant, P. and M, K. (Eds.), Handbook of Research on Holistic Optimization Techniques in the Hospitality, Tourism, and Travel Industry, Business Science Reference, Hershey, PA, USA, Al, A., Ahmed, A., Aljarbouh, A., Donepudi, P., & Choi, M. (2021). *Detecting Fake News using Machine Learning: A Systematic Literature Review.* Retrieved from www.psychologyandeducation.net

Bates, Stephen and Hastie, Trevor and Tibshirani, Robert (2021). *Cross-validation: what does it estimate and how well does it do it?* ArXiv, abs/2104.00673

Barbado, Rodrigo & Araque, Oscar & Iglesias, Carlos. (2019). A framework for fake review detection in online consumer electronics retailers. Information Processing and Management. 56. 10.1016/j.ipm.2019.03.002.

Bentéjac, Candice & Csörgő, Anna & Martínez-Muñoz, Gonzalo. (2019). A Comparative Analysis of XGBoost.

Bottou, Leon; Bousquet, Olivier (2012) *Optimization for Machine Learning: The Trade-Offs of Large-Scale Learning.* United Kingdom: Cambridge MIT Press.

Breiman, L. Random Forests. Machine Learning 45, 5–32 (2001). https://doi.org/10.1023/A:1010933404324

Brownlee, J. (2017). *A Gentle Introduction to the Bag-of-Words Model*. Machine Learning Mastery. https://machinelearningmastery.com/gentle-introduction-bag-words-model/

Cantallops, A.S. and Salvi, F. (2014), "New consumer behavior: A review of research on eWOM and hotels", International Journal of Hospitality Management, Vol. 36

Cone Research. Game changer: cone survey finds 4-out-of-5 consumers reverse purchase decisions based on negative online reviews. Retrieved from: http://www.conecomm.com/contentmgr/showdetails.php/id4008.20

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3

Devlin, Jacob and Chang, Ming-Wei and Lee, Kenton and Toutanova, Kristina (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.* ArXiv, abs/1810.04805

Duan, W., Yu, Y., Cao, Q. and Levy, S. (2016), “Exploring the impact of social media on hotel service performance: A sentimental analysis approach”, Cornell Hospitality Quarterly, Vol. 57 No. 3

Fang, Y., Wang, H., Zhao, L. et al. Dynamic knowledge graph based fake-review detection. Appl Intell 50, 4281–4295 (2020). https://doi.org/10.1007/s10489-020-01761-w

Han, H.J., Mankad, S., Gavirneni, N. and Verma, R. (2016), “What guests really think of your hotel: Text analytics of online customer reviews”, Cornell Hospitality Report, Vol. 16

Jindal, N and Liu, B. (2007) Review spam detection[C]. In: Proceedings of the 2007 International conference on the World Wide Web:1089-1090.

Jindal, N., & Liu, B. (2008). *Opinion Spam and Analysis.* Retrieved from http://money.cnn.com/2006

Khandelwal, R. (2019, December 28). *Word embeddings for NLP*. Medium. Retrieved from https://towardsdatascience.com/word-embeddings-for-nlp-5b72991e01d4

Kumawat, Deepika & Jain, Vinesh. (2015). POS Tagging Approaches: A Comparison. International Journal of Computer Applications. 118. 32-38. 10.5120/20752-3148.

Kwok, L., Xie, K.L. and Richards, T. (2017), “Thematic framework of online review research: A systematic analysis of contemporary literature on seven major hospitality and tourism journals”, International Journal of Contemporary Hospitality Management, Vol. 29 No. 1,

Li, Jiwei & Ott, Myle & Cardie, Claire & Hovy, Eduard. (2014). Towards a General Rule for Identifying Deceptive Opinion Spam. 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014 - Proceedings of the Conference. 1. 1566-1576. 10.3115/v1/P14-1147.

Liang, D. (2022). *Are BERT Features InterBERTible?* KDnuggets. Retrieved from: https://www.kdnuggets.com/2019/02/bert-features-interbertible.html

Lundberg, Scott and Lee, Su-In (2017). *A Unified Approach to Interpreting Model Predictions.* ArXiv, abs/1705.07874

McInnes, L., & Healy, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction.* ArXiv, abs/1802.03426.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (n.d.). *Fake Review Detection: Classification and Analysis of Real and Pseudo Reviews.*

Noone, B.M. and McGuire, K.A. (2014), “Effects of price and user-generated content on consumers’ prepurchase evaluations of variably priced services”, Journal of Hospitality & Tourism Research, Vol. 38 No. 4, pp. 562–581.

P. Soucy and G. W. Mineau, *A simple KNN algorithm for text categorization. Proceedings 2001 IEEE International Conference on Data Mining*, 2001

Pai, A. (2022, June 23). *Pretrained word embeddings: Word embedding NLP.* Analytics Vidhya. Retrieved from: https://www.analyticsvidhya.com/blog/2020/03/pretrained-word-embeddings-nlp/

Pennington, J., Socher, R., & Manning, C. (n.d.). *GloVe: Global Vectors for Word Representation.* Retrieved from http://nlp.standford.edu/projects/glove/

Qaiser, Shahzad & Ali, Ramsha. (2018). Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. International Journal of Computer Applications. 181. 10.5120/ijca2018917395.

Sahin, E., Tang, C., Al-, M., Texas, R., & Antonio, A.-S. (n.d.). *Fake News Detection on Social Media: A Word Embedding-Based Approach.* Retrieved from https://www.researchgate.net/publication/360797462

Sarkar, D. (2019). Feature Engineering for Text Representation. In Text Analytics with Python (pp. 201-273). Apress, Berkeley, CA.

Song, Y. Y., & Lu, Y. (2015). Decision tree methods: applications for classification and prediction. Shanghai archives of psychiatry, https://doi.org/10.11919/j.issn.1002-0829.215044

Thota, A., Tilak, P., Ahluwalia, S., Lohia, N., Ahluwalia, S., & Lohia, N. (2018). *Fake News Detection: A Deep Learning Approach.* Retrieved from https://scholar.smu.edu/datasciencereviewhttp://digitalrepository.smu.edu.Availableat: https://scholar.smu.edu/datasciencereview/vol1/iss3/10

Tida, Vijay Srinivas & Hsu, Sonya. (2022). Universal Spam Detection using Transfer Learning of BERT Model.

Vujovic, Zeljko. (2021). Classification Model Evaluation Metrics. International Journal of Advanced Computer Science and Applications. Volume 12. 599-606. 10.14569/IJACSA.2021.0120670.

Winastwan, R. (2020). *Visualizing Word Embedding with PCA and t-SNE*. Towards Data Science. https://towardsdatascience.com/visualizing-word-embedding-with-pca-and-t-sne-961a692509f5

Woo Gon Kim and Seo Ah Park. (2017), “Social media review rating versus traditional customer satisfaction: Which one has more incremental predictive power in explaining hotel performance?”, International Journal of Contemporary Hospitality Management, Vol. 29 No. 2, pp. 784–802.

Zhao, X. (Roy), Wang, L., Guo, X. and Law, R. (2015), “The influence of online reviews to online hotel booking intentions”, International Journal of Contemporary Hospitality Management, Vol. 27 No. 6

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa