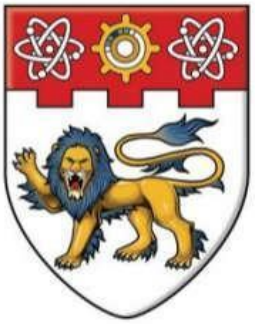


NANYANG TECHNOLOGICAL UNIVERSITY

CZ4034 Information Retrieval

Project Report
Group 14

Matric No.	Name
U1822131H	Sun Yetong
U1722662D	Zhou Hongyu
U1820427G	Liang Ruyin
U1722874L	Zeng Yanxi
U1822704B	Dou Maokang
U1722793H	Li Bingzi

School of Computer Science and Engineering

AY 2020/2021 Semester 2

Crawling
Question 1:
1. How you crawled the corpus (e.g., source, keywords, API, library) and stored it (e.g., whether a record corresponds to a file or a line, meta information like publication date, author name, record ID)
2. What kind of information users might like to retrieve from your crawled corpus (i.e., applications), with sample queries
3. The numbers of records, words, and types (i.e., unique words) in the corpus

Indexing

Question 2:
• Build a simple web interface for the search engine (e.g., Google)
• A simple UI for crawling and incremental indexing of new data would be a bonus (but not compulsory)
• Write five queries, get their results, and measure the speed of the querying

Question 3:
Explain why they are important to solve specific problems, illustrated with examples.

Sentiment Analysis

Question 4: Perform the following tasks:

- Build an evaluation dataset by manually labeling 10% of the collected data (at least 1,000 records) with an inter-annotator agreement of at least 80%
- A simple UI for visualizing classified data would be a bonus (but not compulsory)

- Motivate the choice of your classification approach in relation with the state of the art
- Discuss whether you had to preprocess data (e.g., microtext normalization) and why
- Dataset splits: 7 / 3
- ML grid search (CV) / BERT tuning validation
- multi-task: perform subjectivity & objectivity together
-
- Provide evaluation metrics such as precision, recall, and F-measure and discuss results:
    - Between subjectivity/polarity
    - Between models
    - Between embedding methods
    - Between classical ML / BERT
- Discuss performance metrics, e.g., records classified per second, and scalability of the system

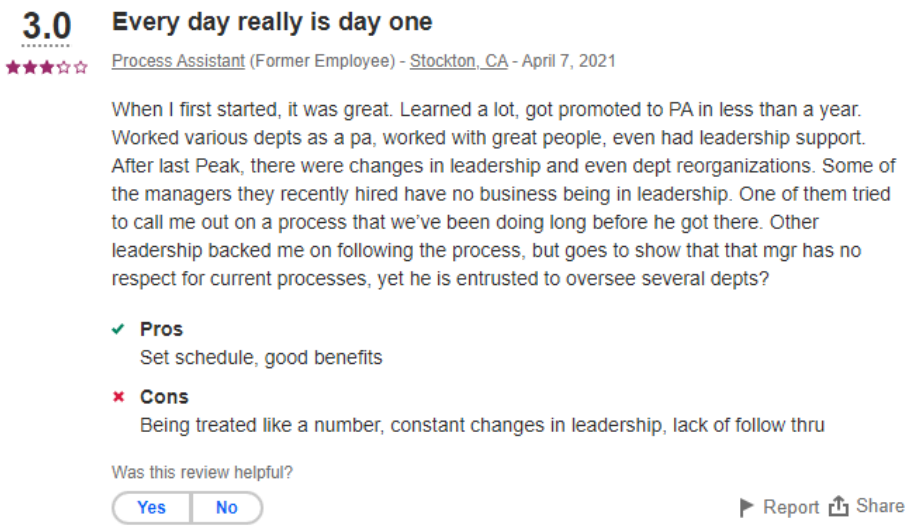
1. Crawling

1.1. How you crawled the corpus and stored it

The data source of our dataset is Indeed website https://www.indeed.com/career/salaries. From this website, we could browse top paying companies by industry. First, we get a list of 143 links of company websites and store it as the csv file named scrape.csv. After we got the list of companies, we went to the company website and according to the observation we could reach the reviews of this company by adding '/reviews' to the end of the previous url. Then we collected the most recent 200 reviews from each company and finally got around 2.8w reviews saved as text file and the according metadata as csv file.

After we collect the list of links of company, we utilize selenium and beautifulsoup to do the web scraping. Beautiful Soup parser HTML into an easy machine readable tree format to extract DOM Elements quickly. It allows extraction of a certain paragraph and table elements with certain HTML ID/Class/XPATH. Selenium is a tool designed to automate Web Browser. It is commonly used by Quality Assurance (QA) engineers to automate their testings Selenium Browser application. Additionally, it is very useful to web scrape because of these automation capabilities: Clicking specific form buttons, inputting information in text fields and extracting the DOM elements for browser HTML code.

The following images show the review example and the according html. We will scrape the title, location, job type and time from each review with the category and assigned IDs as the metadata. Refer to the html code, all those content could be found according to the class name. For review content, after we get the review, we save it as a text file with the id. Each record corresponds to a file. And for metadata file, each record corresponds to a line of the csv file.



The reviews we collect will have more emotional content compared with the news, and the reviews we collect would be more formal than the social platform comment with less emoji and abbreviation.

1.2. What kind of information users might like to retrieve from your crawled corpus (i.e., applications), with sample queries

Users would like to retrieve the reviews from corpus according to different requests.

1.2.1. Search keywords and return matched reviews

sample query: pay

Result would be the reviews that contain the keyword ‘pay’, users only need to retrieve the reviews and check if the keyword contained inside the review.

1.2.2. Company name + aspect

sample query: company name: paypal, query: work life balance

Result would be the reviews from this company and in this aspect. Users need to check the company name from metadata and also check the aspect from the review text file.

1.2.3. Company category + aspect

sample query: company category: Banks and Financial Services, query: culture

Results would be the reviews from companies belonging to the specific category then check for the aspect. Users need to retrieve the category from metadata and then check for the review content of the specific aspect.

1.2.4. Search company location and review keywords, return companies’ reviews based on the given location

sample query: location name: New York, NY, query: benefits

Results would be the reviews from companies in a location then check for the review content if it contains keywords. Users need to retrieve the location from metadata and then check for the review.

1.3. The numbers of records, words, and types (i.e., unique words) in the corpus

total reviews: 28265

=== category Information ===
Number of reviews of each category
Internet and Software: 2017
Industrial Manufacturing: 1764
Health Care: 1638
Banks and Financial Services: 1617
Consulting and Business Services: 1617
Transport and Freight: 1617
Telecommunications: 1596
Retail: 1575
Auto: 1575
Human Resources and Staffing: 1573
Consumer Goods and Services: 1554
Food and Beverages: 1554
restaurants, Travel and Leisure: 1407
Organization: 1365
Construction: 1218
Insurance: 1218
Government: 1176
Aerospace and Defence: 966
Education and Schools: 798
Energy and Utilities: 420

=== Token Information ===
Average token count per review: 62.79766495666018
Total tokens: 1774976
Total types (unique tokens): 38368
=== Sentence Information ===
Average sentence count per summary: 3.926127719794799

2. Indexing and Querying

2.1. Text Preprocessing

After obtaining the crawled reviews from Indeed, the review text needs to be preprocessed such that it removes noises from the text and it is easier for the inverted-index text search engine in the later stage to perform better. In this part, we mainly used a Python library named nltk to preprocess the text data.

2.1.1. Text Tokenization

We used `nltk.word_tokenize()` to tokenize the text. This method transforms a string into an array of substrings. Each substring is a single word or a punctuation.
Example:

Input	Output
`I really enjoy the hiring process. It seems easy. However, if you are an`	`['I', 'really', 'enjoy', 'the', 'hiring', 'process', '.', 'It', 'seems',`

international individual and English is not your first language, then \"GOOD LUCK!!\"	'easy', '.', 'However', ',', 'if', 'you', 'are', 'an', 'international', 'individual', 'and', 'English', 'is', 'not', 'your', 'first', 'language', ',', 'then', '``', 'GOOD', 'LUCK', '!', '!', "''"]

2.1.2. Spelling Correction

There are some spelling mistakes when users input their reviews on Indeed. Spelling check should be performed such that the program could avoid stem misspelled words at the stemming step. Also, Elasticsearch does not have to index misspelled words which decreases querying performance

Isolated word correction is used at this step. The library used to implement this function is called Peter Norvig's spell corrector. This spell corrector helps check each token and finds the most likely correction for the given token.

Sample Input	Output
goverment	government

2.1.3. Removing Stopwords

Removing common words such as "the" and "a" that are appearing in every document reduces file data size. Keeping only meaningful and important words increases efficiency when it comes to querying.
Our list of stopwords is as follows:

```
i, me, my, myself, we, our, ours, ourselves,
you, you're, you've, you'll, you'd, your,
yours, yourself, yourselves, he, him, his,
himself, she, she's, her, hers, herself, it,
it's, its, itself, they, them, their, theirs,
themselves, what, which, who, whom, this,
that, that'll, these, those, am, is, are, was,
were, be, been, being, have, has, had, having,
do, does, did, doing, a, an, the, and, but,
if, of, by, with, then, here, there, d, ll, m,
o, re, ve
```

2.1.4. Stemming

Stemming is a way to reduce a word to its base form by removing its

suffix. By applying stemming to tokens, it is more efficient for queries to retrieve relevant documents with different word forms but the same stem.

PorterStemmer is used in this step and the example is as follows:

Sample Input	Output
social distance	social distanc

2.1.5. Removing Punctuations

Punctuations in reviews carry less meaning as compared to english words. Hence punctuations are removed in preprocessing.

2.2. Indexing

Elasticsearch is used to index the review data, which creates an inverted index to support near-real-time full text search. More specifically, we have indexed the pre-processed review texts (i.e. after tokenization, spell correction, stemming, etc.).

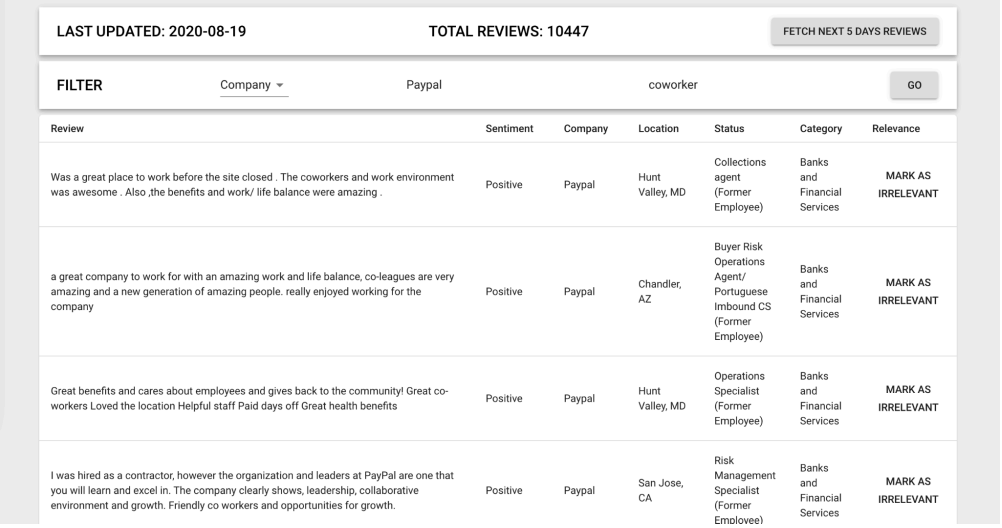
2.3. Simple UI

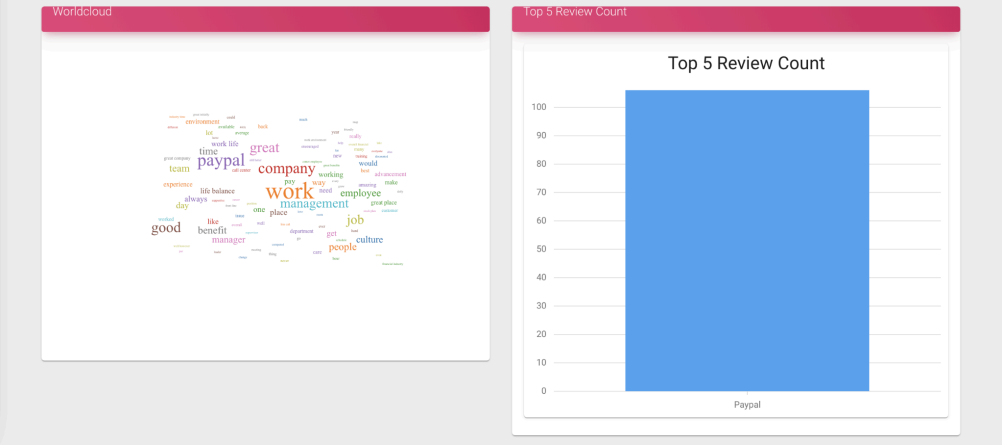
The below snapshot shows the UI of the search engine. For general search, users could input keywords and retrieve relevant reviews.

For other search types such as search by company, location and more, users could utilize the search filter and corresponding input field to obtain query results with conditions.

The UI also allows users to explore the retrieved reviews through different charts.
Bar Chart: Shows percentage of positive, neutral and negative sentiments of retrieved reviews.
Line Chart: Shows the timelines and trends of the sentiment of retrieved reviews.
Histogram: Shows the top 5 companies which contribute most of the relevant reviews.
Word Cloud: Gives users a quick view of review keywords of a given company.
Interactive Search: Marks a review as irrelevant. The query result will be refined and updated

Additionally, the UI supports incremental indexing. Incoming new reviews could be indexed by clicking the button “Fetch next 5 days reviews”. Once clicked, the system will start indexing new reviews and other metadata for the next 5 days. Minutes later, the task will be done and users will see “Last Updated” and “Total Reviews” changes accordingly.





2.4. Querying

2.4.1. Search keywords and return matched reviews

Sample query: pay

Top 3 results

Top	Review
1	This job lies about pay and will change the way they pay you and no pay for down time because there is no hourly pay for fueling pre and post trips even if this takes a while
2	When I started the environment was excellent and pay was excellent. The company has chosen to outsource all decent

	paying jobs and move O&O employees to lower compensation paying jobs.
3	hire you for one job. then have you do another job that is salary and pays way more but doesnt pay you for it. unfair work place that tells you they will pay you a certain amount and then doesnt

speed: 257ms

2.4.2. Company name + aspect

Sample query:
Company name: paypal
Query: wrk life balanc (with misspelled words)

Top 3 results

Top	Review	Company Name
1	Was a great place to work before the site closed . The coworkers and work environment was awesome . Also ,the benefits and work/ life balance were amazing .	Paypal
2	Overall Paypal was a average place to work. The main thing was the work life balance. Two of the perks I found the best, were the sabbaticals and the flex time.	Paypal
3	A great place to work !! Leadership is great good training and the benefits are good. People are friendly and a healthy work environment. Great work/ personal life balance	Paypal

speed: 25ms

2.4.3. Company category + aspect

sample query:
Company category: Banks and Financial Services
Query: culture

Top 3 results

Top	Review	Company Category
1	The company lives and breathes it's corporate culture. If you enjoy a fast pace	Banks and Financial

	work environment with a strong corporate culture then PayPal may be a good company to consider.	Services
2	Great people and culture. Very positive experience. The culture allowed me to try different things and interact with senior management. It is very flat for an organization perspective.	Banks and Financial Services
3	Northwestern Mutual has a culture rooted in helping others attain financial peace and freedom. Everyone is very helpful and gives their time and resources to welcome you into their culture.	Banks and Financial Services

speed: 134ms

2.4.4. Company location + Aspect

Sample query:
Location: new york
Query: career advancement

Top 3 results

Top	Review	location
1	The job itself is easy but can be stressful due to the passengers and the attitudes they give you. Its a great place to make advances and promotions and the hours were great. Although the management is questionable its a great place to expand your career with the U.S Government.	New York State
2	Worked at a NY branch from years indicated above. Not sure what the culture is like now, but my experience at Garda was mostly appalling. Trucks were ill-maintained (at best), days were long, and pay was mediocre. Began to feel like a beast of burden after a while. Pretty dispiriting place. Still have nightmares sometimes about my time there. Would only work here again if there was literally no other option. That said, some people really seemed to enjoy working at Garda (usually, in my experience, ex-military and retired cop types).Not many advancement opportunities when I was there.Moving on from Garda was one of the best decisions I ever made.	New York
3	Although I was here for a short period of time, I	New York

	was able to meet great people and learned a lot. Challenging environment, advanced technology tools.	State

speed: 15ms

2.4.5. Employment status + Company name + query

Company: paypal
Employment status: current employee
Query: advancement

Top 3 results

Top	Review	Status
1	great benefits and opportunities to advance . great perks . Provide tools and resources needed to be succesful . Management provide great feedback needed to achieve monthly goals.	Collector III (Current Employee)
2	PayPal is good company to work for with many career advancement opportunities. The pay and benefits offered are great! I've never worked for a company that has such great benefits from day one. This company truly cares for their employees.	Customer Solutions Teammate (Current Employee)
3	There is limited advancement opportunities. The company and my organization had weak leadership based on nepotism and favoritism. There are better places to seek employment.	Merchant Sales/ Business Development (Former Employee)

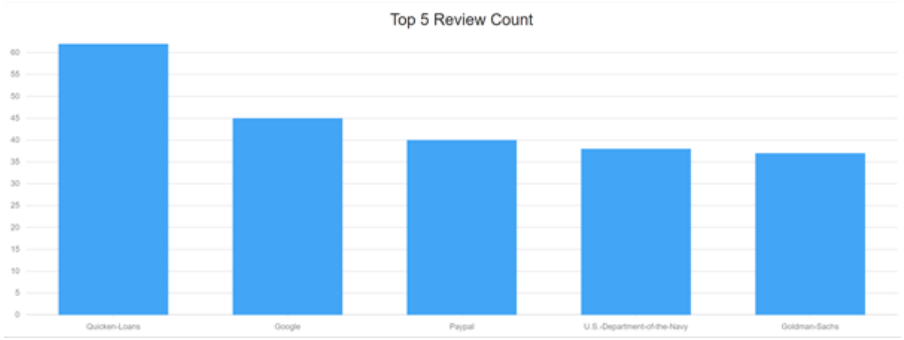
speed: 12ms

3. Innovations for Enhancing Indexing and Ranking

3.1. Enhanced Search

Histogram

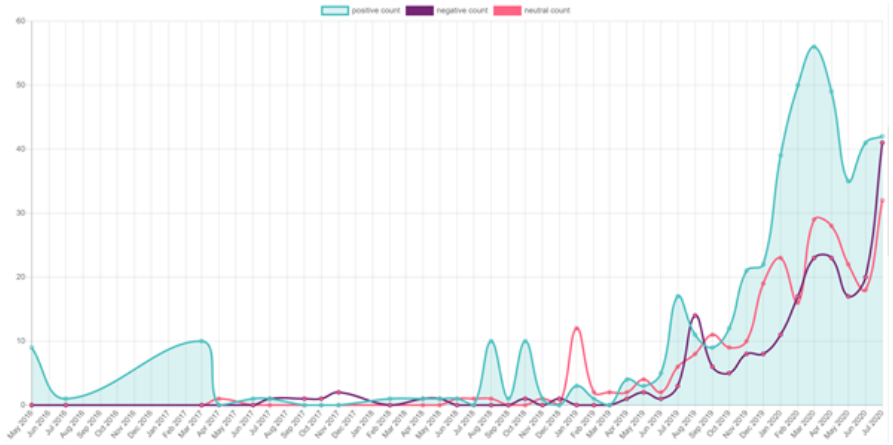
When we want to search for "work life balance" using the search bar, only showing the search result of reviews provides less information about which companies have better "work life balance". A histogram shows the number of reviews mentioning "work life balance" of each company could help users have a further understanding on the related reviews.



Line Chart

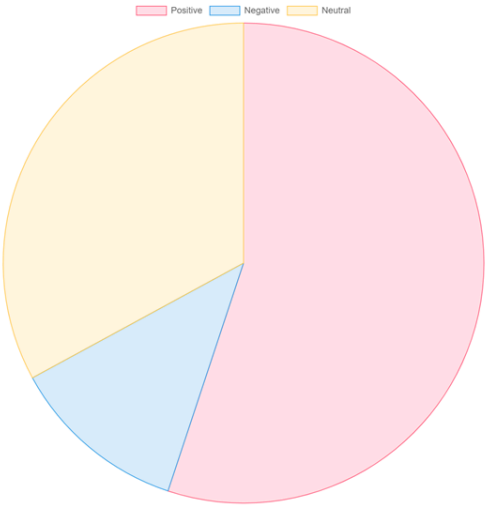
Companies may have different performance at different time quantum. The line chart helps users to have a better visualization on the company's performance.

E.g. User input “Morgan Stanley” at the company search bar, the line chart data shows positive, negative and neutral reviews counts of “Morgan Stanley” at different time periods. The reviews trend can be easily observed through this line chart.



Pie Chart

E.g., Users may want to search reviews for a specific company such as “Paypal”. A pie chart is used to demonstrate the percentage of positive, negative and neutral reviews more clearly.



Table

A table is used in order to show more information about the reviews.

E.g. User input “work life balance” into the search bar, in addition to reviewing content, users may also want to see information like “Location”, “Company”, “Sentiment” for better understanding of the data.

Top	Review	Sentiment	Company	Location	Status	Category
1	great benefits and opportunities to advance . great perks . Provide tools and resources needed to be succesful . Management provide great feedback needed to achieve monthly goals.	Positive	Paypal	Hunt Valley, MD	Collector III (Current Employee)	Banks and Financial Services
2	PayPal is good company to work for with many career advancement opportunities. The pay and benefits offered are great! I've never worked for a company that has such great benefits from day one. This company truly cares for their employees.	Positive	Paypal	La Vista, NE	Customer Solutions Teammate (Current Employee)	Banks and Financial Services
3	There is limited advancement opportunities. The company and my organization had weak leadership based on nepotism and favoritism. There are better places to seek employment.	Neutral	Paypal	Omaha, NE	Merchant Sales/ Business Development (Former Employee)	Banks and Financial Services

Word Cloud

The Word Cloud could provide a better visualization of mentioned words of a company.

E.g. User input “Paypal” into the company search bar, mentioned words in reviews of “Paypal” should be returned and appear at the wordcloud. The weight of each word should appear upon moving the mouse over. Thus, the most weighted word for this example is “work” and the second weighted word is “paypal” etc.

3.2. Interactive Search

The system allows users to tag a review as “not relevant” such that the query results will be refined based on the relevance feedback.

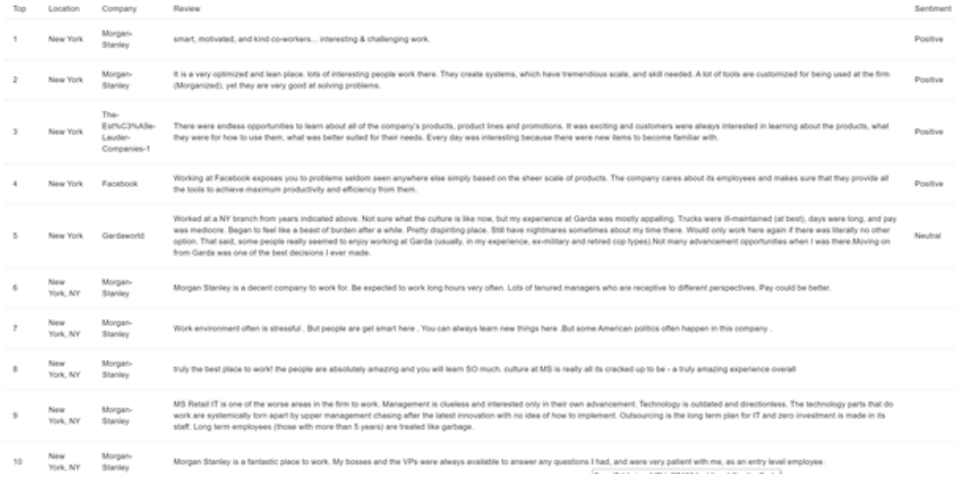
To refine the query results based on users’ relevance feedback, we have dug into a classic and well-known technique called the Rocchio algorithm. Lucene’s MoreLikeThis class is a variant of the Rocchio algorithm, which is used as the underlying function of ElasticSearch’s “more_like_this” query.

When a review is marked as “irrelevant” by the user, the system uses 5 shown reviews (which are not marked as “irrelevant”) to feed into the Rocchio algorithm, thus getting updated query results which will be of higher relevance to the user’s intended search results.

3.3. Geospatial Search

The system allows users to search for reviews by location.

E.g., If the user lives in “New York” and he may like to search for top reviews in “New York”. So only reviews for which are located at “New York” would be returned. This helps users easier filter the data regarding the location.



Top	Location	Company	Review	Sentiment
1	New York	Morgan-Stanley	smart, motivated, and kind co-workers... interesting & challenging work.	Positive
2	New York	Morgan-Stanley	It is a very optimized and lean place. lots of interesting people work there. They create systems, which have tremendious scale, and skill needed. A lot of tools are customized for being used at the firm (Morganized), yet they are very good at solving problems.	Positive
3	New York	The-E4%C3%A9e-Leader-Companies-1	There were endless opportunities to learn about all of the company's products, product lines and promotions. It was exciting and customers were always interested in learning about the products, what they were for how to use them, what was better suited for their needs. Every day was interesting because there were new items to become familiar with.	Positive
4	New York	Facebook	Working at Facebook exposes you to problems seldom seen anywhere else simply based on the sheer scale of products. The company cares about its employees and makes sure that they provide all the tools to achieve maximum productivity and efficiency from them.	Positive
5	New York	Gardaworld	Worked at a NY branch from years indicated above. Not sure what the culture is like now, but my experience at Garda was mostly appalling. Trucks were ill-maintained (at best), days were long, and pay was mediocre. Began to feel like a beast of burden after a while. Pretty dispiriting place. Still have nightmares sometimes about my time there. Would only work here again if there was literally no other option. That said, some people really seemed to enjoy working at Garda (usually, in my experience, ex-military and retired cop types).Not many advancement opportunities when I was there.Moving on from Garda was one of the best decisions I ever made.	Neutral
6	New York, NY	Morgan-Stanley	Morgan Stanley is a decent company to work for. Be expected to work long hours very often. Lots of tenured managers who are receptive to different perspectives. Pay could be better.	
7	New York, NY	Morgan-Stanley	Work environment often is stressful . But people are get smart here . You can always learn new things here .But some American politics often happen in this company .	
8	New York, NY	Morgan-Stanley	truly the best place to work! the people are absolutely amazing and you will learn SO much. culture at MS is really all its cracked up to be - a truly amazing experience overall!	
9	New York, NY	Morgan-Stanley	MS Retail IT is one of the worse areas in the firm to work. Management is clueless and interested only in their own advancement. Technology is outdated and directionless. The technology parts that do work are systemically torn apart by upper management chasing after the latest innovation with no idea of how to implement. Outsourcing is the long term plan for IT and zero investment is made in its staff. Long term employees (those with more than 5 years) are treated like garbage.	
10	New York, NY	Morgan-Stanley	Morgan Stanley is a fantastic place to work. My bosses and the VPs were always available to answer any questions I had, and were very patient with me, as an entry level employee.	

4. Machine Learning Approach

4.1. Evaluation Dataset

4.1.1. Labelling Process

1500 reviews were randomly sampled from the data collected for manual labeling. In order to achieve a balanced evaluation dataset with at least 80% inter-annotator agreement, the six group members were splitted into 3 pairs of 2, and each pair worked on the same 500 reviews based on a common labeling standards (Annex). The reviews were labeled as “Neutral”, “Positive” or “Negative”.

After the 1500 reviews were labeled by the annotators, the reviews that had inconsistent labels were collected to undergo a second round of labelling by another annotator who has not seen this review before. By majority voting, the labels were determined in the second round.

As a result, we have obtained a roughly balanced and consistently-labeled dataset with the label counts stated below:

```
Positive    516
Neutral     500
Negative    471
```

Note that for subjectivity detection, 250 reviews from each of the “Positive” and “Negative” classes were sampled combined to form a “Non-neutral” class with 500 reviews, together with the 500 reviews of “Neutral” class.
For the polarity detection task, reviews from the “Positive” and “Negative” were used.

4.1.2. Splitting Data for Training, Validation and Testing

In view of the need to conduct both subjectivity detection, polarity detection and multitask prediction in the later sections, the reviews from each class were randomly splitted into training and testing data at a ratio of 8:2. Therefore, we have obtained a balanced training set and balanced testing set to be shared across all models. In addition, some models require a validation set (e.g. the BERT model). For such models, the training set was further splitted into a training set and a validation set at a ratio of 8:2.

4.2. Data Pre-processing

4.2.1. Balanced data in training set

	neutral (0)	positive (1)	negative (2)
original instance	390	422	377
subjectivity detection	390	390	
polarity detection	NA	422	377

As the table shows, the data instance of objective and subjective are sampled to be the same to avoid bias. In polarity detection, the number of instances are not significantly different and the dataset is small. Thus, all the labeled instances are kept.

4.2.2. text cleaning

The data crawled from the website contains unreadable characters due to the formatting of the webpage.

	text	label
617	Best company in usa.good benefits. I recomende this company for be the best at all thing and aspecs..was good sperience.....and we have layoff beacuse move to paso..texas..was realy good company in florida.	1
686	Itâ€ s a delivery job. Depends on the neighborhoods that you are delivering to. They are helpful and usually flexible with hours and schedule when needed.	0
837	Know who you are working for. The diversity spill is bull. Management mostly non-blacks. Something to get by if you donâ€ t care about semantics.	0
580	Educational, Positive Work Environment, Great Training, Management is fair, the most enjoy able part of the job is at will, company really cares for employees.	1
195	I currently work here itâ€ s not bad I just hate that Iâ€ m timed in the register and we get hours based on how good we are.You basically have to be fast in the floor and in the register mostly when the main managers are there or they will really stress you out about going faster which will make you have anxiety.	1

In sentiment analysis, numbers are not important. Therefore, the first step of pre-processing is to clean the text data so that it only contains alphabets.

The step is done by applying the regular expression operations.

	text	label
617	best company in usagood benefits i recomende this company for be the best at all thing and aspecswas good sperienceand we have layoff beacuse move to pasotexaswas realy good company in florida	1
686	its a delivery job depends on the neighborhoods that you are delivering to they are helpful and usually flexible with hours and schedule when needed	0
837	know who you are working for the diversity spill is bull management mostly nonblacks something to get by if you dont care about semantics	0
580	educational positive work environment great training management is fair the most enjoy able part of the job is at will company really cares for employees	1
195	i currently work here its not bad i just hate that im timed in the register and we get hours based on how good we areyou basically have to be fast in the floor and in the register mostly when the main managers are there or they will really stress you out about going faster which will make you have anxiety	1

4.2.3. Stop Words

Stop words are frequent words that appear in a language, such as “and, the, is” in English. They are not important in sentiment analysis, or other tasks like summarization and information extraction. In this project, the stopwords set from NLTK corpus is used.

4.2.4. Stemming

The goal of stemming is to reduce a word to its base form by crudely chopping off the end of the word by rule-based approach. The most widely used model is Porter Stemmer.

4.2.5. Lemmatization

Lemmatization is similar to Stemming. However, lemmatization reduces the word to its base form on morphological level.

original text after alphabet cleaning	remove stopwords	stemmed text	lemmatized text	remove stopwords & stemmed	remove stopwords & lemmatized
its a delivery job depends on the neighborhoods that you are delivering to they are helpful and usually flexible with hours and schedule when needed	delivery job depends neighborhoods delivering helpful usually flexible hours schedule needed	it a deliveri job depend on the neighborhood that you are deliv to they are help and usual flexibl with hour and schedul when need	it a delivery job depends on the neighborhood that you are delivering to they are helpful and usually flexible with hour and schedule when needed	deliveri job depend neighborhood deliv help usual flexibl hour schedul need	delivery job depends neighborhood delivering helpful usually flexible hour schedule needed

4.2.6. Comparison between Different Approaches

The model used here is SVM. The embedding method used is Bag of Words. The results are compared and analysed which text-preprocessing are better.

Text pre-process	Test Accurac	F1 Score	Recall Score	Precision

ing	y			
original text after alphabet cleaning	0.66	0.66	0.66	0.66
remove stopwords	0.58	0.60	0.62	0.58
stemmed text	0.65	0.65	0.64	0.65
lemmatized text	0.66	0.67	0.67	0.66
remove stopwords & stemmed	0.60	0.60	0.58	0.61
remove stopwords & lemmatized	0.59	0.60	0.61	0.58

Table: SVM

From the table above, it can be observed that in in SVM model, original text ≅ lemmatized text > stemmed text. Comparing the results listed here and results of other models, cleaned original text is selected as the text pre-processing method which will be used throughout the experiment. The reasons are listed below:

- Comparing results of the models, training results using original text is relatively good and stable.
- Clearly, stopwords removal is not suitable for sentiment analysis of comments. The reason is that stopwords removal can remove common words that express the sentiment, like “not”.
- Stemming and lemmatization has the similar problem as stopwords removal. They are good pre-processing methods when performing tasks like concept extraction. However, they can change a word’s meaning which is disastrous to sentiment analysis. For example, stemming can convert a negative word like worthless to a neutral/positive word worth.

4.3. Models

4.3.1. Embedding Methods

4.3.1.1. Bag of Words

BoW is a commonly-used feature extraction method to represent the

occurrence of words in a chunk of text. The steps performed by BoW model to get embeddings of a text is:

- List all words in the text to form a corpus
- Convert text to corpus length vectors, the number represents the occurrence of a word in this chunk of text

Advantages:

- Easy to understand and implement
- Widely applied in text classification problems

Disadvantages:

- Does not preserve the contextual meaning

4.3.1.2. Doc2Vec

Doc2Vec is an algorithm which aims to produce vectors of a document. The Doc2Vec model is based on the Word2Vec model. It uses an unsupervised learning method to embed a paragraph to a user-defined fixed-length vector. The algorithm Distributed Memory Model of Paragraph Vectors (PV-DM) is inspired by the continuous bag-of-words (CBOW) algorithm in Word2Vec. PV-DM produces document vectors along word vectors from the text data, concatenating them to obtain the gradient by backpropagation. Another approach Distributed Bag of Words version of Paragraph Vector (PV-DBOW) is similar to skip-gram approach in Word2Vec.

Advantages:

- Preserves the contextual meaning

Disadvantages:

- Takes time to get vectors during prediction time
- Performance degrades for small text size

4.3.1.3. BERT

4.3.2. Choice of models

4.3.2.1. BERT

BERT (Bidirectional Encoder Representations from Transformers) is a state-of-the-art Transformer-based neural network approach for natural language processing tasks developed by Google. It is pre-trained to obtain bidirectional representations for a text from both its left and right context. The pre-trained BERT model can be fine-tuned for a wide range of NLP tasks, including sentiment analysis, question answering and summarisation with outstanding performance.

There are a few techniques that made BERT powerful:

- Bidirectional context: it is designed to take both the previous and following text into consideration for a more comprehensive understanding of the context.
- Transformers mechanism: It reads entire sequences of tokens at once instead of reading sequentially. The attention mechanism allows for learning contextual relations between words.
- (Pre-trained) contextualized word embeddings - As introduced

in the ELMO (embedding from language model), a vector assigned to a token/word is a function of the entire sentence containing that word. Therefore, the same word can have different word vectors under different contexts.

In our tasks, the publicly available pre-trained BERT model was used. Along with a BERT tokenizer, our review data was fed to finetune the BERT model for the sentiment analysis task: subjectivity and polarity detections.

4.3.2.2. Naive Bayes

Naive Bayes classifiers are a group of simple probabilistic classifiers based on Bayes theorem. The model used in this project is Multinomial Naive Bayes classifier, which is suitable for classifying discrete features, such as the word count in the Bag of Words model. Therefore it is likely to be suitable for text classification tasks.

4.3.2.3. SVM

Support vector machine is a supervised learning algorithm which constructs hyper-planes in a dimensional space to achieve goals like classification. In the preprocessing section, the text data has been converted to vectors in a multi-dimensional space. SVM is a reasonable model to get good results.
The model used in this project is C-support Vector Classification (SVC). SVC, NuSVC and LinearSVC in sklearn are wrappers of the same algorithm in libsvm. SVC in sklearn supports multiple kernels, including linear and polynomials. While selecting parameters, grid search is applied. Linear kernel is selected.

4.3.2.4. XGBoost

Boosting is an algorithm which adds new models to existing ones to improve the ensembled model performance. Extreme Gradient Boosting (XGBoost) is a type of gradient boosted decision tree. The advantage of XGB is that it is very fast. It is applied to perform regression or classification on tabular dataset.The disadvantage of XGBoost is that it has a high variance approach and overfits on small dataset.

4.3.3. Grid Search

Grid search is a method to tune the hyper parameters of an estimator. The results are determined by Cross Validation score. There are two grid search methods available: Exhaustive Grid Search and Randomized Parameter Optimization. The difference between them is Exhaustive Grid Search will search the entire hyper parameter combinations while Randomized Parameter Optimization samples from these combinations. Since the data and model are not complex, Exhaustive Grid Search is used in this project with 5-fold Cross Validation. For example, apply grid search on SVM model:

hyper parameter set	Cs = [0.001, 0.01, 0.1, 1, 10] gammas = [0.05, 0.1, 0.15, 0.20, 0.25] degrees = [0, 1, 2, 3, 4, 5, 6] kernels = ['rbf', 'linear', 'poly']
selected parameter	{'C': 0.01, 'degree': 0, 'gamma': 0.05, 'kernel': 'linear'}

4.4. Results

4.4.1. Subjectivity Detection

Embedding methods	Classifier	Test Accuracy	F1 Score	Recall Score	Precision	Number of records classified per second
BOW	SVM	0.668	0.660	0.645	0.676	1804
BOW	Naive Bayes	0.641	0.626	0.600	0.653	8313
BOW	XGB	0.627	0.613	0.591	0.637	5595
Doc2 Vec	SVM	0.610	0.636	0.682	0.595	132
Doc2 Vec	XGB	0.618	0.641	0.682	0.605	128
BERT	BERT	0.735	0.735	0.736	0.742	62

4.4.2. Polarity Detection

Embedding methods	Classifier	Test Accuracy	F1 Score	Recall Score	Precision	Number of records classified per second
BOW	Naive Bayes	0.926	0.922	0.883	0.965	9955
BOW	SVM	0.915	0.916	0.926	0.906	2681
BOW	XGB	0.872	0.874	0.883	0.865	5139
Doc2Vec	SVM	0.851	0.854	0.872	0.837	131
Doc2Vec	XGB	0.835	0.832	0.819	0.846	141
BERT	BERT	0.943	0.945	0.944	0.955	56

4.5. Discussion

4.5.1. Embedding Methods

BERT has the best performance in subjectivity detection. BERT is one of the state of the art models in NLP. Compared to BoW embedding, BERT considers both the previous and the following context. Thus, BERT understands the contextual meaning of words in a chunk of text, which is a huge advantage for sentiment analysis. Thus BERT gives better results than machine learning models using BoW embedding.

The surprising finding is that BoW embedding outperforms Doc2Vec embedding. The main reason is our dataset is too small. Doc2Vec works well only with large volumes of data. Doc2Vec is an advanced embedding method and usually trains on up to millions instances of data in published papers. The number of words per document is much larger than our dataset as well.

4.5.2. Models

BERT outperforms all the machine learning models. Comparing Multinomial Naive Bayes, SVM and XGB, MNB and SVM has similar results and XGB has the worst result.

In most cases, SVM should work better than MNB in text classification. However, the performance of SVM and MNB depends on the nature of text. MNB could perform better than SVM on classifying short documents, which is the case in the project.

XGB is a popular and effective classification model. However, the performance of XGB is poor in this project. The reason is again related to the size of data samples. When training on a small amount of data, the overfitting problem of XGB is much bigger than other tree-based models due to its complexity. For instance, Naive Bayes performs better because it has a simpler hypothesis function, which results in less overfitting.

4.5.3. Subjectivity and Polarity

The best result is 73.5% for subjectivity detection and 94.3% for polarity detection, both from BERT. The reason for low accuracy in subjectivity detection and very high accuracy for polarity detection comes from the dataset itself. Unlike news reports, the comments from the job searching website usually contain sentiments. In the training data, there are rarely completely neutral instances. When labelling the data, comments which contain both positive and negative sentiments are marked neutral. It is even difficult for humans to determine whether they are neutral, biased towards positive or negative. This results in the low accuracy for subjectivity detection.

4.5.4. Classifier Efficiency

The testing data needs to be embedded before classification, which counts towards the classification time. Comparing the number of records classified per second for different embeddings, the speed of BoW >> Doc2vec > BERT. BERT has the best result but significantly slower. In the future, the system needs to consider the efficiency and accuracy trade-off between BoW and BERT. Comparing the speed of different machine learning models, the speed of MNB > XGB > SVM. Unlike the efficiency gap between BoW+machine

learning and BERT, they are at the same order of magnitudes.
In a word, to maximize the efficiency, BoW+MNB should be selected. To maximize the accuracy, BERT should be selected. Among all these models, BERT is selected as the model for the search engine at the current stage since the size of the dataset is small.

5. Enhanced Sentiment Classification

In this section, we will explore techniques for enhancing the sentiment classification. We have experimented with multitask classification and ensemble classification with sentiment analysis, which will be discussed below.

Note that the models used in the two tasks are different (details elaborated below), the two enhancement techniquess were not used jointly and their results would be reported separately.

5.1. Multitask Classification

Multitask classification is an approach that performs multiple subtasks jointly. In our study, we combined the subjectivity detection task and polarity detection task together, to examine the improvement of using one model to accomplish the two tasks jointly.

5.1.1. Why important + example

It is important to combine the tasks of subjectivity detection and polarity detection, as in most use cases these two tasks need to be conducted together for sentiment analysis. Since we are not sure about whether our review data carriers polarised sentiment, it could be meaningless to directly apply polarity detection on data that may carry neutral content. Hence, the classifier needs to handle the detection of neutral content and additionally the polarity of sentiment if applicable.

This process could be done by using a single multitask model or a pipeline of two models handling subjectivity and polarity detections separately. We therefore would like to experiment whether one model dealing with the two tasks could enhance our classification.

For example, for a review like :

“If you can commit to the long hours the pay is great and the company culture is amazing. They also train you on everything needed even if you have little mortgage banking experience. ”

Before extracting its sentiment, it is safer to first classify this review into neutral/non-neutral. However, using a pipeline of subjectivity and polarity detection could result in lower classification efficiency as two models are involved. Therefore, a multitask model is needed to handle such situation.

5.1.2. Methods

In view of the outstanding performance of the BERT model in both subjectivity and polarity detections, it is chosen as the model for the multitask classification. In order to know whether a review data is neutral and which polarity of sentiment it carries, we used a training and testing dataset with three labels: Neutral, Positive and Negative.

To evaluate the performance of the multitask model, we have constructed a prediction pipeline with the best performing subjectivity detection model and polarity detection model. In both cases, BERT model performs the best. As shown below, the test review data would be first fed into the subjectivity classifier. If the review is non-neutral, it would next be fed into the polarity classifier to determine whether its sentiment is positive or negative.

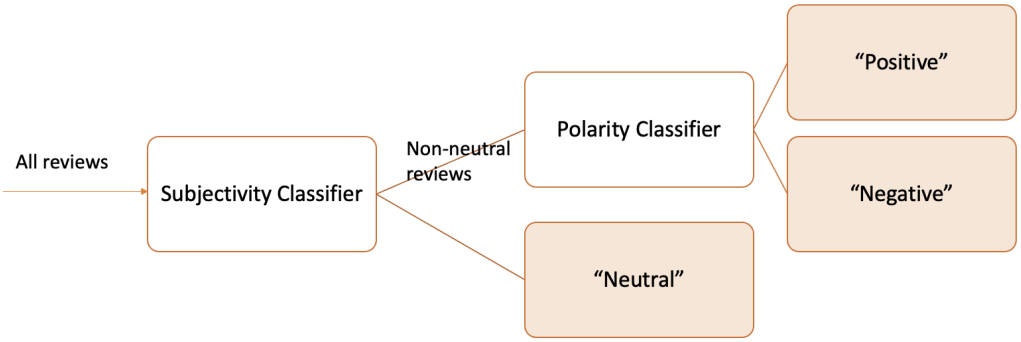


figure 1. Pipeline for subjectivity detection polarity detection, to be compared against the multitask model

5.1.3. Results

The test classification accuracies and evaluation metrics obtained from the single multitask model and the pipeline are stated below.

	Test Accuracy	F1 score	Recall Score	Precision
Single Multitask Model	0.76	0.76	0.76	0.76
Pipeline of Two Sub-models	0.661	0.663	0.661	0.667

Table 1. Test accuracies, F1 scores, recall scores and precisions for the single multitask model and the pipeline of two sub-models

It is observed that the single multitask model outperforms the pipeline in all four evaluation metrics. In addition, we have examined the evaluation metrics for the three classes respectively, as shown below.

Single Multitask Model

	precision	recall	f1-score
neutral	0.629	0.667	0.647
positive	0.787	0.794	0.791
negative	0.871	0.804	0.836

Pipeline of Two Sub-models

	precision	recall	f1-score
neutral	0.570	0.627	0.597
positive	0.717	0.702	0.710
negative	0.729	0.660	0.693

Table 2. F1 scores, recall scores and precisions for the single multitask model and the pipeline of two sub-models for “Neutral”, “Positive” and “Negative” respectively.

From Table 2, it is observed that the single multitask model has outperformed the pipeline of two sub-models in classification of all three classes. In both models, classification of the “Neutral” class has the lowest performance, and the classification of “Positive” and “Negative” classes share similar performance in the two models respectively.

5.1.4. Discussion

It is interesting to note the difference in performances between the two approaches. Given the performances of the subjectivity detection (test accuracy = 0.73) and for polarity detection (test accuracy = 0.94), the multitask model (test accuracy = 0.76) has shown an improved performance as compared to the subjectivity detection while showing a degraded performance as compared to the polarity detection. However, the pipeline (test accuracy = 0.66) has shown a degraded performance even compared to the subjectivity detection.

This could be explained by the different ways of how the multitask model and pipeline model make decisions. In the multitask model, the class label is determined by comparing the probabilities of them belonging to each class. Therefore, all three classes were compared at once without a certain sequence. However, in the pipeline approach, the accuracy for subjectivity detection is the key bottleneck as the first task followed by polarity detection. Unlike a single multitask model where the overall accuracy could be enhanced by considering probabilities for three classes all together, the pipeline approach accumulates the errors from each stage which results in a degraded performance. Hence, we conclude that a multitask model outperforms a

pipeline approach with two binary sub-models.

5.2. Ensemble Classification

Ensemble classification is an approach that considers predictions from multiple base models to enhance classification. In our study, we have chosen three classical machine learning models and two ensemble techniques and evaluated their impact on subjectivity detection and polarity detection respectively.

5.2.1. Why important + example

Ensemble classification is widely adopted to improve model performance from multiple base models. Given the fact that the performance of the machine learning models for the subjectivity and polarity detections might have been restricted by limited sample size, and the quality of text collected (as there is limited pure "neutral" text, and some samples from "neutral" class contain a mixture of opinionated content).

For example, for a review with complex sentiment expression like:

"I love QL. My VP is amazing. You get what you put in. They try to make it fun. I miss being in an office. I deal with some angry clients but remember they.are not made at you. Don't pass the buck, solve it and improve their experience. "

The base classifiers learnt to predict this review differently:

Naive Bayes	SVM	XGBoost
Neutral	Non-neutral	Neutral

Therefore, by combining their predictions to learn the group-truth, the base classifiers could complement each other's strengths and weaknesses if they are independent models and not correlated with each other.

5.2.2. Methods

For ensemble learning, predictions from multiple classifiers were taken into consideration to produce the final prediction outcome. In our study, the individual predictions on the common test set from the three classical machine learning models used in the previous section, namely Naive Bayes, SVM and XGBoost, were taken as inputs to the ensemble model.

Note that the BERT model is not used in this section because it overfitted on the training data due to the limited sample size. Since it produces a training accuracy of nearly 1.0, the ensemble model will assign 100% weight on it and hence fail to learn. Therefore, we only consider the classical machine learning model in this section to

study the potential improvement brought by ensemble classification.

5.2.2.1. Majority Voting

As the most basic ensemble technique used, majority voting takes the most agreed label among predictions from all classifiers and assigns it to an instance. It helps to reduce the error rate if the classifiers all perform better than random guessing and are independent of each other.

5.2.2.2. Stacked Ensemble

Stacked ensemble makes use of a meta-classifier and multiple base-classifiers. It takes the prediction outputs from the base-classifiers as inputs to train a meta-classifier that maps these inputs to the ground-truth. In our study, three models with grid search were used as the meta-classifier to find the best performing one: logistic regression, XGBoost and random forest.

5.2.3. Results

Accuracy of each ensemble technique	Majority Voting	Stacked - Logistic Regression	Stacked - Random Forest	Stacked - XGBoost
Subjectivity Detection	0.650	0.641	0.641	0.641
Polarity Detection	0.931	0.931	0.910	0.910

Table 3. Accuracies of ensemble classifiers using majority voting and stacking (with meta-classifiers of logistic regression, random forest and XGBoost respectively) for subjectivity detection and polarity detection

Accuracy of each base classifier	Naive Bayes	SVM	XGBoost
Subjectivity Detection	0.641	0.668	0.627
Polarity Detection	0.926	0.915	0.872

Table 4. Accuracies of single classifiers for subjectivity detection and polarity detection, whose prediction outputs are used as inputs for the ensemble classifiers

As compared to single models, the ensemble classifiers have shown average performance in subjectivity detection and improved performance in polarity detection. Among the ensemble techniques, majority voting appears to be optimal. For subjectivity detection, majority voting has shown to be better than two base models (naive bayes and XGB) and a slightly degraded performance than SVM. For polarity detection, majority voting has shown a better performance than all three base models.

5.2.4. Discussion

The results have shown that, in certain use cases, appropriate use of ensemble techniques is able to improve performance of sentiment classification. For polarity detection, it is able to learn to combine the results so as to complement the weaknesses of the base models.

On the other hand, we also noticed that ensemble classification has limited improvement in the subjectivity detection task. By examining the predictions made by each base model, we found that the outputs from the three base models are highly similar, thereby also explaining the similar performance of all three stacked ensemble techniques. It is possible that the quality of the data under neutral label has hindered the model to learn, which also led to three highly correlated base models and consequently an average ensemble classification performance.