Bachelor Thesis

Analysing Student Comments on RateMyProfessors.com Using NLP Techniques

Submitted by

GIACOMO GIOLI

Matriculation Nr.: 17678541

On

26th May 2020

At

ZURICH UNIVERSITY OF APPLIED SCIENCES

School of Management and Law

Faculty of Business IT

Supervised by

DR. BLEDAR FAZLIJA

Institute of Wealth & Asset Management

Management summary

The assessment of teaching methods and faculty performance is an important step enabling educational institutions to continuously improve their teaching methods and study offers. Typically, schools conduct internal surveys to assess their performance. In most cases, however, the results of these surveys are not disclosed to the public. Therefore, several online platforms have emerged, which allow students to evaluate their teachers publicly. The most popular online evaluation platform, RateMyProfessors.com, currently features over 15 million evaluations covering more than 1.8 million teachers. So, how can schools use this large amount of publicly available data to generate useful insights?

In order to answer this research question, a dataset containing 1,637,435 evaluations for 134,375 teachers from 605 schools selected with a random approach using web scraping techniques was built. Intermediate questions were defined in order to answer the research question, such as whether it is possible to use computational techniques to distinguish good from bad teachers based on the language used by the students. The individual questions were elaborated and answered using theoretical knowledge and statistical models.

Using natural language processing and machine learning techniques it was demonstrated that it is possible to distinguish positive evaluations from negative evaluations, easy subjects from difficult subjects as well as good teachers from bad teachers with accuracies of over 90%. Furthermore, thanks to the correlations discovered between the quality of teaching as perceived by students, the level of difficulty as perceived by students and the helpfulness of the teacher, it was possible to predict the quality of teaching and the level of difficulty based on the students language. Finally, it was demonstrated that, using statistical models, it is possible to identify topics concerning the faculty performance and teaching methods in evaluations of online courses.

Although random approaches to data collection have been chosen to allow the results to be generalized, this cannot be considered universally valid, as the platform from which the data was extracted offers the possibility to evaluate only institutes in the United States, Canada and the United Kingdom. It is therefore necessary to consider possible differences in the way teachers in other cultures are evaluated.

In conclusion, natural language processing and machine learning techniques can

be applied for the analysis of online evaluations. Schools can therefore use these techniques to generate useful information about their teachers and their teachers' performance based on online evaluations. This approach, however, should not be looked at by schools as an alternative to the typical evaluation activity, but as an extension, allowing them to analyze aspects not normally considered in typical school evaluations.

Contents

List of Figures

List of Tables

1 Introduction

The assessment of teaching methods and teachers by students is an important step that enables schools to continuously improve their teaching methods and study offers (Marsh, 1987, p. 259). Schools typically conduct internal surveys, asking their students to evaluate the subjects and teaching methods applied by the teachers. However, the results of these surveys may not be made publicly available or only partially with the students who have attended the classes (Azab, Mihalcea & Abernethy, 2016, p. 438). Due to the non-disclosure of the results of these surveys, the emergence of many online platforms encouraging the evaluation of schools and teachers was inevitable. Among these platforms, the most popular is RateMyProfessors.com[1] (RMP). On RMP, students can evaluate their school and teachers in an anonymous way. Students have the opportunity to present their assessment by evaluating different aspects of the school and teachers. For teacher evaluation, these aspects include clarity, helpfulness and difficulty. While assessing the campus, students can present their assessment taking into account the following aspects: reputation of the school, services offered, happiness, location, quality of food served, social activities, opportunities and campus security. In addition to both teacher evaluation and school assessment, students can provide an open comment. RateMyProfessors.com currently features over 15 million ratings for over 1.8 million professors in the United States, Canada and the United Kingdom (RateMyProfessors, 2020). Thanks to the information published on these online platforms, students can make more thoughtful decisions about their academic journey (e.g. which school to enroll in or which courses to attend) (M. J. Brown, Baillie & Fraser, 2009, p. 91). Previous research has found strong evidence that publicly available information about the reputation of teachers and schools influences students' educational career decisions (C. L. Brown & Kosovich, 2015).

However, it may be argued that due to the fact that online ratings can be entered by anyone at any time, they may be biased by several factors and therefore may not reflect student learning and faculty performance (Otto, Sanford & Ross, 2008). In fact, previous studies have found that online evaluations may be influenced by teacher's personality, charisma, physical appearance and grading leniency (Felton, Mitchell & Stinson, 2003; Otto et al., 2008). Consequently, online evaluations may

[1] www.ratemyprofessors.com

be influenced by emotions and therefore may present a halo effect. The halo effect shows that students who make evaluations give either a high or a low rating without providing detailed feedback on various aspects of faculty performance (Felton et al., 2003). Furthermore, due to the fact that students decide of their own will to evaluate their teachers and their institute, it can be argued that evaluations on their own are affected by self-selection bias. In contrast, other studies suggest that it is possible that online evaluations may not be biased in general. In fact, even if some students provide biased evaluations, they may be balanced between other evaluations that do not necessarily have a bias (Otto et al., 2008).

Aim of the thesis The objective of the thesis is to analyze students' comments on RMP using natural language processing (NLP) techniques to investigate language patterns that can be used to generate useful insights about teachers and teaching methods. So, the research question of the thesis is: how can schools use publicly available data about their teachers to generate useful information? The question is elaborated using the following questions which help to develop the discussion and results:

- Is it possible to distinguish positive evaluations from negative evaluations?
- Is it possible to predict the quality score expressed by the student in the rating?
- Is it possible to distinguish easy from difficult subjects?
- Is it possible to predict the level of difficulty as perceived by students?
- Is it possible to distinguish between online courses and classroom courses?

In addition:

- Can all student comments be used to distinguish good teachers from bad teachers?
- Is it possible to predict the teacher's overall score using all student comments?

And finally:

- Is it possible to recognize specific topics that are discussed in the evaluations of online courses?

Web scraping[2] and reverse engineering techniques are used in order to gather the necessary data to build the dataset.

2 Theory

In order to be able to answer the research question in a clear way, as a first step it is necessary to acquire knowledge about studies that have already been carried out in the field of Natural Language Processing (NLP) and automated text classification. Furthermore, it is necessary to understand the different concepts that revolve around NLP and ML. The following points will serve to acquire knowledge about the current state of research and to clarify the fundamental concepts of the thesis.

2.1 Definitions

Machine Learning is a subcategory of the artificial intelligence concept. Machine Learning is a science that deals primarily with the development of efficient and accurate algorithms for prediction purposes. These computational methods use experience to make accurate predictions. Experience means the past information that is made available to the algorithm. Machine learning is divided into three main tasks: classification, regression and clustering (Mohri, Rostamizadeh & Talwalkar, 2012, p. 2-3).

The classification problem deals with assigning a category to each item. For example, a classification algorithm could classify newspaper articles into the categories politics, economics, sports and technology based on the content of the article. In order for the algorithm to learn how to classify articles into the correct categories, it is necessary to provide experience to the algorithm to learn from. This means that a large number of newspaper articles would have to be collected and manually assign a category to each article (Mohri et al., 2012, p. 3).

The regression problem is about predicting a value for each item. A concrete example for a regression algorithm is predicting the market value of a real estate property. In order to solve the problem, the algorithm should be able to have experience of properties, their characteristics and market value available to it. For this reason it is necessary to train the algorithm to recognize which are the characteristics

[2]Web scraping: a computer technique used to extract data from a website

of the property that increase or decrease the value of the property (Mohri et al., 2012, p. 3).

The clustering problem deals with identifying and partitioning groups within large amounts of data. Clustering related tasks are mainly unsupervised. This means that the algorithm tries to learn and perform the task almost completely autonomously. An example of the clustering task is the analysis of traffic on the telephone network to identify different groups and different behaviors within the network (Mohri et al., 2012, p. 3).

Natural Language Processing is an area of research and application that deals with understanding how computers can be used to understand and manipulate natural language. NLP is based on the foundations of various disciplines such as mathematics, statistics, linguistics and artificial intelligence. Thanks to the interdisciplinarity of this research field, NLP deals with problems such as automated language translation, speech recognition, text classification and automated information extraction (Chowdhury, 2005, p. 51).

Web Scraping, also known as screen scraping, data mining or web harvesting, is an automated technique of collecting data from websites. Web scraping allows to extract data from websites quickly and automatically without the need for human intervention. Web pages are normally constructed with HTML code that composes the graphic elements of the page. With web scraping techniques it is possible to extract the information contained in the graphic elements that are displayed in the web browser (Mitchell, 2018, p. 9).

2.2 State of research

When it comes to analysis and statistics, people almost always think only of numbers and functions. However, in an increasingly digitalized world, the amount of data generated becomes larger and larger and the variety of this data becomes greater and greater. In this case, therefore, we are talking about a large quantity and a large variety of data. Data can be structured (e.g. databases and tables), semi-structured (such as XML[3] and JSON[4]) or unstructured (text, images, video and

[3]XML is a metalanguage for the definition of markup languages.

[4]JSON is a format used for the exchange of data between client/server applications.

audio). Structured and semi-structured data are very easy to analyze by a computer as they have a structure, whether it be strong or weak. Structured and semi-structured data have metadata that often describe their structure and the relationships between them. However, a large part of the content that is generated online by users and within companies is not structured (see Facebook, Instagram, company reports, blogs, etc.). In fact, it has been estimated in a study conducted at IBM that around 80% of the world's data is unstructured (Schneider, 2016). A large part of the value of data is stored in text, audio, images and videos. Therefore, the analysis of unstructured data represents both a challenge and a potential for extracting valuable information (Manning & Schütze, 1999). Economically, it was estimated in a study by Gualtieri and Yuhanna (2016) that companies on average use only between 27% and 40% of their data to generate useful insights for the company. This means that companies have enormous unexploited potential. In fact, if companies were able to analyze also unstructured data they own (and the data freely available on the web) they could greatly increase the value of the insights generated. To exploit this untapped potential, classic statistical and analytical techniques are not sufficient. Therefore, new techniques to extract information from unstructured data are necessary.

Research in natural language processing (NLP), a subfield of linguistics and computer science began in the 1940s when there was a need to decode enemy ciphers during World War II (Liddy, 2001). At that time, the term machine translation (MT) rather than the term NLP was used.

MT models used the ideas from cryptography and language translation theory. However, MT was based on a very trivial approach. As a matter of fact, it was mainly a dictionary lookup and rearrangement of the words in the order required by the target language (Liddy, 2001). Initially, the results produced by MT were of poor quality, and it was soon realized that this task was much more complex than expected. In the post-war period, other areas of NLP research, such as speech recognition, began to emerge. Thanks to the continuous development of new language theories and new parsing algorithms, progress was made in the research of this area. In the 1950s, models were expected within just a few years that would be capable of producing results comparable to those of a human being (Liddy, 2001). However, the researchers of the time were once again wrong.

Due to the excessive enthusiasm on the part of the researchers and consequently the little progress made in this area, the funds allocated to NLP research were drastically

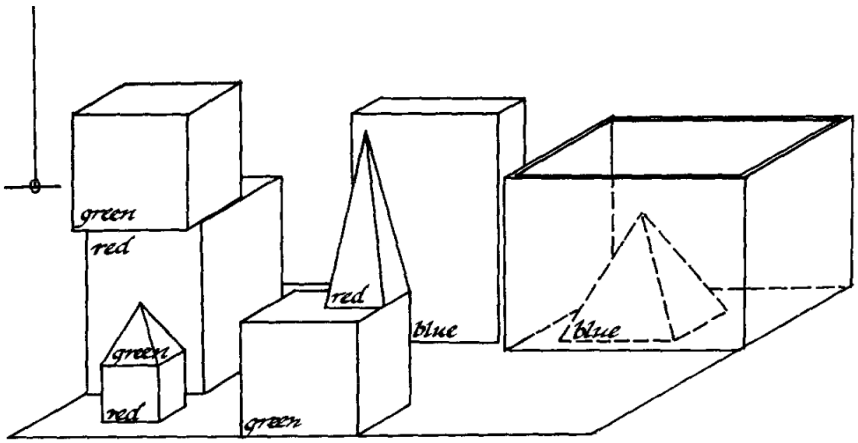

Figure 1: SHRDLU's interface (Winograd, 1972, p. 8)

reduced. For this reason, applied research in this field faced a slowdown (Liddy, 2001). However, theoretical research in the field of linguistics was able to continue. Despite the slowdown in research, at the beginning of the 1970s, thanks to a program created by Winograd (1972), it was demonstrated that computers could understand and interpret natural language. In the program created by Winograd called SHRDLU, a three-dimensional space was shown in which some three-dimensional shapes with different characteristics and colours were arranged. Some objects shared the same characteristics. The goal of the program was to capture the user's requests by writing them in the command line and then perform the required operation. The requests could be for example "put the green pyramid in the box" or "move the blue parallelepiped over the green cube". SHRDLU had to be able to distinguish the various geometric figures based on the characteristics described by the user and then perform the required operation. The difficulty in the task was in recognising the object the user intended, because on the three-dimensional plane there were objects that shared the same characteristics. The program, although limited in its functionality, proved that a computer could understand natural language. Figure 1 on page 6 shows the interface designed by Winograd.

In the 1990s the NLP research field started to grow rapidly again thanks to several factors such as the availability of better performing computers to a wider audience and the availability of large quantities of digital text. At the end of the 20th century, research in the field of NLP mainly used a linguistic approach based on the rules of language and theoretical foundations. With the arrival of the Internet and thus digitalisation, large amounts of electronic data began to be generated. With it

also came the concept of big data. Using large datasets and thanks to the gains in computing power due to technological developments, research in the field of NLP shifted to a data-based approach (Connolly et al., 2016). Despite the change of approach, hybrid approaches are still often used to produce robust systems capable of generating high quality results. A hybrid approach makes it possible to create systems that are able to interpret with great accuracy thanks to the rule-based approach and that are robust in input variation thanks to the data-based approach.

Thanks to the progress made in the field of NLP more and more subfields of research were created, and with these also a growing number of practical applications. The most widely used practical applications of NLP are the following: information retrieval, information extraction, question answering, summarization, machine translation, text classification and named entity recognition. NLP techniques can be applied virtually anywhere, as natural language is used in any context where there is interaction with a human being. Examples of a widely used NLP-based applications are google translate and DeepL. A practical example of text classification is that applied by email providers, which automatically sorts incoming mail into the user's various folders basing on the email's content. However, text classification techniques can be applied in many other areas, as it is one of the most widely used applications of NLP.

Thanks to the extensive activity within the NLP research community, several open-source libraries have been developed to support the creation of new NLP applications. These include NLTK[5], SpaCy[6], SciKit[7] and AllenNLP[8]. These libraries provide a set of features that allow users to perform operations on text data, such as lemmatisation, stemming ans TF-IDF vectorisation .

So far only one study that tried to classify good and bad teachers listed on RateMyProfessors.com with NLP techniques has been identified. However, the study conducted by Azab et al. (2016) was limited to simple binary classifications (good and bad). Although their study was limited to a binary classification, it showed that it is possible to distinguish the quality of teaching with very high accuracy using NLP techniques. Based on the knowledge generated and described in Azab et al. (2016) study, this thesis aims to go beyond a binary classification, trying to predict

[5] nltk.org
[6] spacy.io
[7] scikit-learn.org
[8] allennlp.org

the quality perceived by students by mixing different NLP and ML techniques.

2.3 Hypotheses

This part of the thesis focuses on hypotheses that have been developed based on previous research. We present the hypotheses with their rationale, which will be either accepted or rejected on the basis of the results provided by the analysis.

Hypothesis 1a: NLP and ML techniques can be used to predict the quality of teaching perceived by students based on the textual comments of the evaluations.

Hypothesis 1b: NLP and ML techniques can be used to predict the level of difficulty perceived by students based on the textual comments of the evaluations.

Rationale: In their study, Azab et al. (2016) proved to be able to distinguish a good teacher from a bad teacher with an accuracy of about 90% using only the comments provided by the students. However, in their study, Azab et al. (2016) distinguished only between a good and a bad teacher, without estimating the grade given by the student. Theoretically, since the grade given by the student is a whole number between [1, 5], instead of two classes (good and bad) the comments could be classified into five classes (1 to 5). Probably, having more classes in which to classify comments, it will be necessary to fine-tune the model further in order to avoid false classifications and consequently a decrease in accuracy.

Hypothesis 2: By including the helpfulness score in the analysis it is possible to predict more accurately the quality of teaching perceived by students.

Rationale: In a study conducted by Otto, Sanford and Wagner (2005) it was shown that there is a correlation between the quality of teaching perceived by students and the teacher's helpfulness. In fact, following logical reasoning one can come to the conclusion that a teacher who makes his time available to the students, the students perceive a higher quality of teaching. On the other hand, if a teacher is not available to the students, the students perceive a lower quality of teaching. Therefore, it is expected that by exploiting this correlation, it is possible to better predict the quality of teaching perceived by students.

Hypothesis 3: By including the difficulty score in the analysis it is possible to predict more accurately the quality of teaching perceived by students.

Rationale: In a study by Otto et al. (2005), the various correlations between the variables reported in teaching evaluations on RMP were researched. In the study it was shown that the difficulty of the subject is positively correlated with the quality of the teacher's teaching. Since one of the questions we want to answer is whether it is possible to predict the quality of teaching perceived by students, we expect that by exploiting the correlation between the difficulty of the subject and the quality of teaching, it is possible to predict more accurately the quality perceived by students.

Hypothesis 4: If the teacher's gender is included in the analysis of the students' comments, then it is possible to predict more accurately the teacher's overall score.

Rationale: Mengel, Sauermann and Zölitz (2019) have demonstrated the existence of a gender bias in teaching evaluations. In fact, the findings of Mengel et al. (2019) showed that female teachers generally receive lower evaluations than their male colleagues. This bias may be due to stereotypes such as the questioning of female instructors' competences or the lack of confidence perceived by students in female teachers (Mengel et al., 2019, p. 28). The bias described by Mengel et al. (2019) may be leveraged to adjust the weights of the model to increase its accuracy.

Hypothesis 5: NLP and ML techniques can be used to distinguish between online and classroom courses, based on the textual comment of the evaluation.

Rationale: If it is possible to distinguish between good and bad professors using purely the language of students' comments, as demonstrated in the study of Azab et al. (2016), then we suppose that it is also possible to distinguish online courses from classroom courses by interpreting the language used by students.

Hypothesis 6: It is possible to recognise specific topics discussed in the comments of the evaluations concerning online courses using NLP techniques.

Rationale: If the online evaluations are not biased by the halo effect described in the study by Otto et al. (2008), then it is possible to recognise topics concerning

the performance of the faculty that are discussed in the evaluations. However, if due to the self-selection of samples and personality, charisma and physical appearance of the teacher the online assessments are biased by the halo effect, then it should not be possible to detect distinct aspects of faculty performance.

3 Data harvesting

The collection and preliminary analysis of data is perhaps the most important step in the Machine Learning process. In fact, data preparation represents 80% of the work in the ML process (Wilder-James, 2016). The gathering of the necessary data is the step that allows to start the whole process. Indeed without data it would not be possible to create any model. Whereas the preliminary analysis of the collected data allows to comprehend in detail the data that will be processed and used for the creation of the model. A detailed knowledge of the dataset is necessary to define which aspects of the data can be leveraged to create a robust and accurate model. In addition, a detailed preliminary analysis of the collected data allows the discovery of any missing or dirty data that could create problems in the subsequent steps. For instance, it has been estimated by IBM that poor data quality costs the US economy around $3 trillion each year (Redman, 2016). For this reason it is important to collect data in an orderly manner and carry out a preliminary analysis to find any missing or dirty data. This chapter discusses data extraction techniques and proposes a preliminary statistical analysis of the collected data.

3.1 Data extraction

In order to create a ML model that is able to predict the quality of teaching perceived by students based on their textual comments, it is necessary to compile a dataset to train and validate the model. As schools often do not disclose the results of internally conducted teaching quality surveys, the data available on RateMyProfessors.com is used. The legal aspects of extracting data from websites and the techniques used to extract data from RateMyProfessors.com are discussed in the following points. Furthermore, in this chapter a statistical analysis of the dataset is proposed in order to better understand how to proceed with data processing.

3.1.1 Legal aspects

Before collecting data from RateMyProfessors.com it is necessary to consider the legal aspects related to the extraction of data from third party websites. Web scraping is often considered a data collection technique of dubious legality. Indeed, over the past few years, large companies such as LinkedIn and Ryanair have filed lawsuits against companies that have extracted data from their websites (Kernel, 2019). The case filed by Ryanair against Expedia was settled between the two parties without the intervention of the court of appeal. Expedia agreed to remove Ryanair's flight information from its website (Schaal, 2019). On the contrary, in the decree no. 17-16783 issued by the U.S. Court of Appeal regarding LinkedIn v. HiQ Labs, the judge decided that LinkedIn could not deny HiQ to collect information about users with a public profile on the professional social network. The decision taken by the U.S. Court of Appeals represented a very important point in the era of data and privacy regulations.

From these two cases the legal conditions that define the boundary between legality and illegality of web scraping techniques can be deduced. In Ryanair's case, the data extracted by Expedia was used for commercial purposes, so it must be inferred that scraping data from third party websites for commercial purposes was and remains illegal. Scraping of copyrighted content, such as videos on YouTube, is also illegal. However, it is theoretically legal to extract other information from YouTube, such as video titles and user comments. Whereas, in the case of LinkedIn it can be deduced that it is illegal to collect information from websites that require registration and authentication. Theoretically, for websites that do not require authentication, it is completely legal to extract data, as these sites cannot require the user to accept any terms of service before the user sees the data. However, if the scraping algorithm causes, in any way, problems for the website, such as server congestion and consequently creates connection problems for other users, the company may still take legal action. In any case, for companies that do not want third-party algorithms to collect data from their website, there are countermeasures to prevent it. For example, it is possible to use a robots.txt file on the web server to block requests not coming from a web browser. Alternatively, it is possible to use CAPTCHA technology to request human verification to proceed to the website.

Therefore it is possible to summarize the legal conditions that make scraping

legal or illegal in the following points:

1. Data extracted from third party websites may not be used for commercial purposes unless the website owner specifically consents.

2. It is strictly forbidden to extract data from websites that require authentication.

3. Extracting content subject to copyright is also prohibited.

4. The scraping algorithms must work in such a way that the functionality of the website and the web server is not compromised in any way, otherwise the website owner may take legal action.

To understand whether the extraction of data from RateMyProfessors.com is criminally punishable, a small analysis has been made using the listed conditions. Condition number 1 is met as the data is extracted for educational purposes and purely for research purposes. The RateMyProfessors.com website does not require any kind of authentication to display the data, so the second condition is also fulfilled. For the third condition the discussion becomes a bit more complicated. In fact, point 4 of the Terms of Use Agreement[9] on RateMyProfessors.com states that the content of the website, such as design, text, images and illustrations are subject to copyright. However comments and reviews made by users are considered User Generated Content (UGC) and therefore are not property of RateMyProfessors.com. Although user comments are still subject to copyright, they are used for research purposes only and are under no circumstances redistributed or republished anywhere else. Therefore condition number 3 is also satisfied. The data that will be extracted from RMP will only be textual and numerical data, therefore of small dimensions. Accordingly it is assumed that the functionality of the website and the web server will not be compromised, so condition number 4 is also respected.

3.1.2 The dataset

The school rankings are often biased and do not represent a real ranking. In fact, there are multiple rankings that try to rank the various schools (see Times Higher Education, QS World University Ranking, etc.), and in each of these rankings, the same order is rarely found. The reason for this may be that different aspects are

[9] Terms of Use Agreement RateMyProfessors.com

used to assess different schools or simply because the ones who draw up the rankings favors the schools in their own country (Holmes, 2018). For this reason, unlike the approach chosen by Azab et al. (2016), a random approach has been chosen. This means that instead of selecting universities manually based on a ranking, universities are selected randomly using an algorithm. This detaches us from the rankings and creates an equal probability of selecting any school. Furthermore, by selecting universities randomly, there is theoretically a higher probability of collecting more evenly distributed data. Finally, the lack of bias allows us to generalize the results and apply them to the broader frame (Horton, 2019).

The scraping process was launched several times to avoid generating excessive traffic on the RMP servers and, therefore, to comply with condition number 4 described in section 3.1.1. The scraping algorithm requires a list of all universities registered on RMP, then randomly selects a number of them. Once the schools have been selected, the algorithm requires the list of all teachers for each of the schools. Finally, the algorithm collects all the evaluations for each teacher of each university. Since the scraping algorithm is launched more than once, it is possible that a school is randomly selected more than once, and with it also the teachers of that school and the teachers' evaluations. For this reason, duplicates are deleted from the three datasets. Table 1 on page 13 reports the number of records in each dataset.

	Schools	Professors	Evaluations
Number of records	605	134,375	1,637,435

Table 1: Number of records for each dataset

Professors dataset The teacher dataset is composed of the following characteristics: the name of the teacher, the unique ID assigned by RMP, the department where the professor teaches, the name and ID of the school where the instructor teaches, the average of the helpfulness scores assigned by the students, the average of the difficulty scores assigned by the students, the average of the clarity scores assigned by the students and the general score which represents the average of the availability and clarity scores. Table 2 on page 14 shows a section of the professors dataset.

pk_id	teacherfullname_s	teacherdepartment_s	schoolname_s	school_id	averagehelpfulscore_rf	averageeasyscore_rf	averageclarityscore_rf	averageratingscore_rf
1364631	Brian Zack	Languages	Princeton University	780	4.8	2.3	4.5	4.7
136441	Elizabeth Bogan	Economics	Princeton University	780	4.0	3.2	4.0	4.0
556516	Rober George	Political Science	Princeton University	780	4.54	4.17	4.71	4.62
243005	Paul Krugman	Economic	Princeton University	780	3.12	2.94	3.18	3.15

Table 2: Section of professors dataset

Evaluations dataset In addition to the professor's name and the name of the class, there are many other attributes in the evaluation dataset, such as the number of thumbs up and thumbs down given by other users to the evaluation. The number of thumbs up and thumbs down shows the opinion of other users on what is stated in the evaluation. The evaluation also shows whether the student would take the course again and some tags that the student may associate with the teacher. In addition, the student has the option to report in the assessment the note received in the subject. Furthermore, the dataset contains the attribute *isForOnlineClass* which reports if the student has followed the course online or in class. Additionally each grade has an attribute that describes how the textbook was. This attribute ranges between 0 and 5, where the value 0 represents non-use and the number 5 an extensive use. Finally, each evaluation has scores for clarity, helpfulness and difficulty. These attributes take an integer value between 1 and 5. To complete the evaluation, a textual comment provided by the student is also reported. Table 3 on page 15 gives an overview of the ratings dataset.

class	department	teacherfullname_s	mandatory	grade	isForOnlineClass	ratingsTags	textbookUse	thumbsUp	thumbsDown	wouldTakeAgain	clarityRating	difficultyRating	helpfulRating	comment
ENG111	Languages	Brian Zack	non mandatory	Audit/No Grade	True	Gives good feedback– Respected– Amazing lectures	5	1	0	True	5	4	5	He is so nice. He does not matter when you come to class.. just listen. Useful classes. Thank you
ECON101	Economics	Elizabeth Bogan	non mandatory	B-	True	Gives good feedback– Respected– Amazing lectures	5	1	1	True	1	2	3	Professor Bogan is a hero among women. She is the successful lady economist of your dreams. She is a brilliant lecturer and patient and fair in office hours. Take all of her classes, fantastic for Pub...
POL305	Political Sciences	Rober George	non mandatory	A	False	Gives good feedback– Respected– Inspirational	0	1	2	True	4	5	3	Brilliant lecturer. Engages students in a large lecture. Examines both sides to arguments and leaves his personal politics outside the classroom. A wonderful experience.
CONINTERP	Economics	Paul Krugman	non mandatory	A+	True	Gives good feedback– Inspirational– ACCESSIBLE OUTSIDE CLASS	5	1	0	True	3	3	5	Wasn't a tough class but Krugman is surprisingly boring to listen. Hard to believe he is the #1 economist in the world.

Table 3: Section of ratings dataset

3.2 Preliminary analysis

To better understand the content of the two datasets a preliminary analysis of the data is performed.

Percentage of evaluations without comment Since the focus of this paper is mainly on the use of NLP, we want to find out what percentage of the evaluations do not have a textual commentary. In addition to evaluations without comment, evaluations with minimal commentary, such as "OK", "...", ":)", etc. are also considered. The analysis showed that only about 0.21% of the evaluations have no comment.

Percentage of evaluations with grade The grade received by the student in the subject may influence the teacher's assessment. For this reason we want to know what percentage of the assessments include the student's grade. The analysis shows that only 6 out of 25 assessments (24%) report the grade. Although the percentage is not very high, this translates into about 150,000 assessments.

Correlations To understand which features are correlated with others, correlations are searched. Figure 2 shows a heatmap describing the correlations between the variables in the evaluation dataset. The analysis also confirms the hypotheses of the study conducted by Otto et al. (2005). Moreover, it can be seen from figure 2 that there are several other correlations between the variables of the dataset. The variable *wouldTakeAgain* is strongly positively correlated with the variables *clarityRating* and *helpfulRating*, while the same shows a moderate negative correlation with the feature *difficultyRating*. The same type of correlation can be seen between the variable *grade_numerical* and *clarityRating*, *difficultyRating* and *helpfulRating*. The *grade_numerical* variable has been computed to transform the grade from the American format ($A\pm-F$) to a number from 0 to 12, where the number 0 represents the F and the number 12 the grade A+. It is also possible to see that there is a positive correlation between the grade and the *wouldTakeAgain* feature.

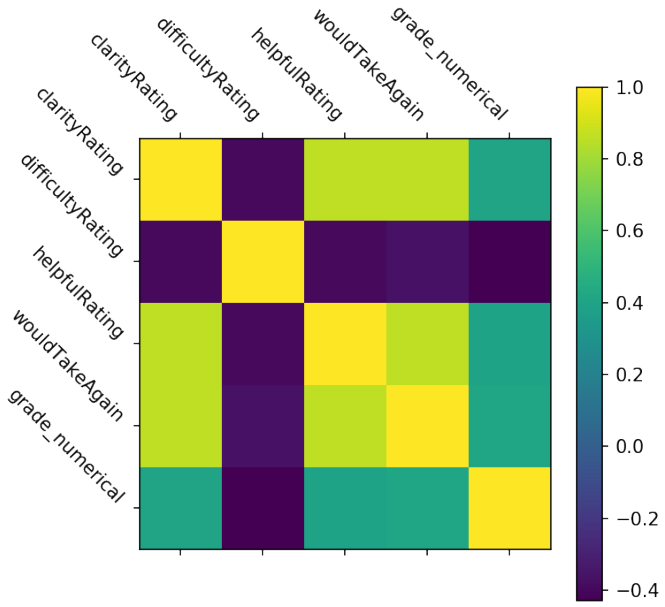

Figure 2: Correlations in evaluations dataset

Distribution Since one of the questions to be answered is whether it is possible to predict the scores given by the students based on their comments, we are interested in viewing the distribution of the various scores. Figure 3 shows the distributions for the variables *clarityRating*, *helpfulRating* and *difficultyRating*. The clarity and helpfulness scores show a very similar but very imbalanced distribution. In fact, the clarity and helpfulness scores tend upwards. In contrast, difficulty scores show a typical normal distribution.

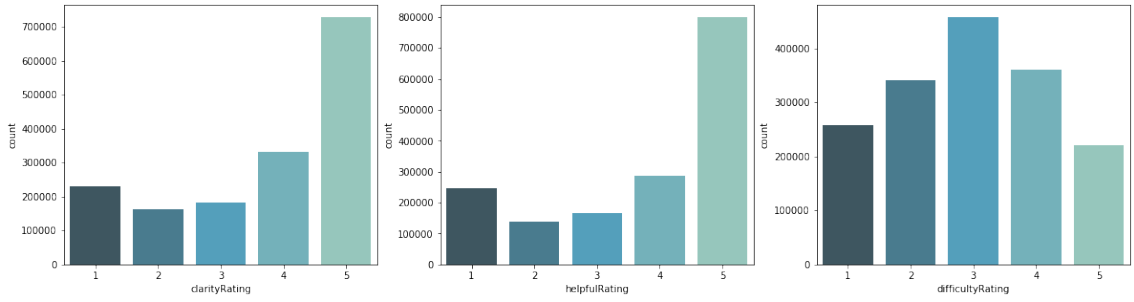

Figure 3: Distributions clarity, helpfulness & difficulty scores

Furthermore, in order to analyze hypothesis number 4 concerning the existence of a bias on the gender of the teacher, we are interested in the gender distribution of teachers. However, RMP does not specify whether the teacher is male or female and therefore this data is not even present in our dataset. For this reason we use the open-source library called gender-guesser to determine the gender of the teacher

based on the first name.

Figure 4 shows the gender distribution of teachers. As can be seen, in most cases it has been possible to derive gender from the first name. However, there is still a good part where it has not been possible to derive gender from the first name. This may be due to the fact that there are teachers who have an exotic name that is not present in the database. There is a small part labelled as*andy* which means that the teacher's name is androgynous (both male and female).

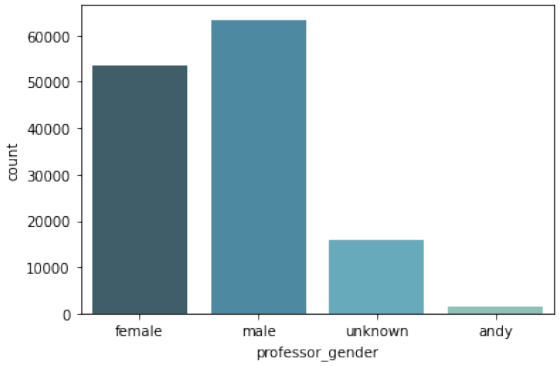

Figure 4: Distributions of professor gender

Moreover, in order to better understand the content of the dataset we are interested in visualizing which are the teaching departments with the most teachers and which departments are the most evaluated. Figure 5 shows the number of teachers per department, while Figure 6 shows the number of assessments per department. Not surprisingly, both graphs show the same departments, as the number of assessments is also influenced by the number of teachers. Interestingly, the order of departments in the two graphs is not the same. This shows that students in certain departments are more likely to share their opinion about teachers than others.

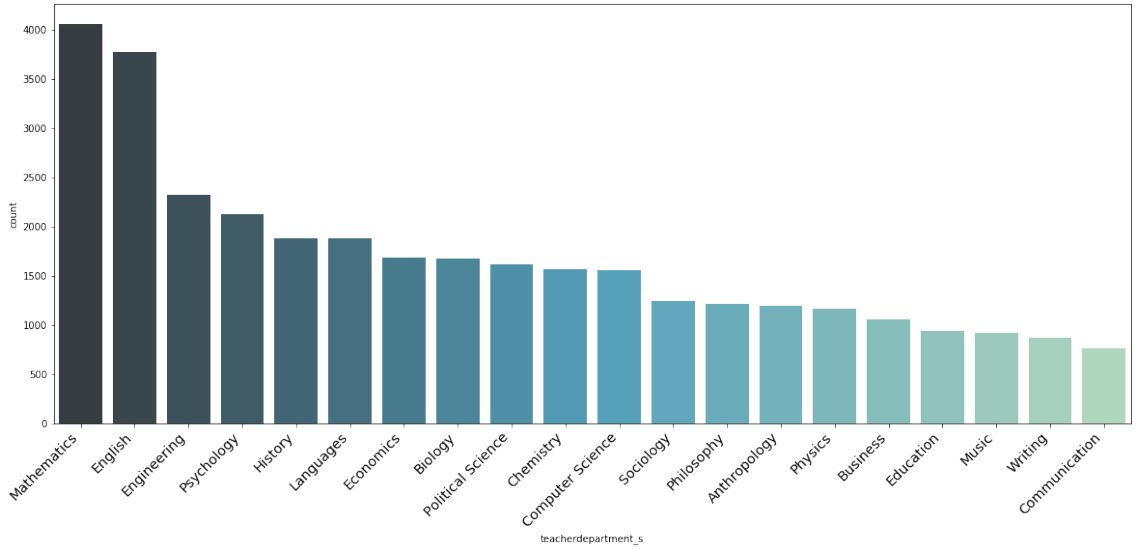

Figure 5: Distribution professors by department (top 20)

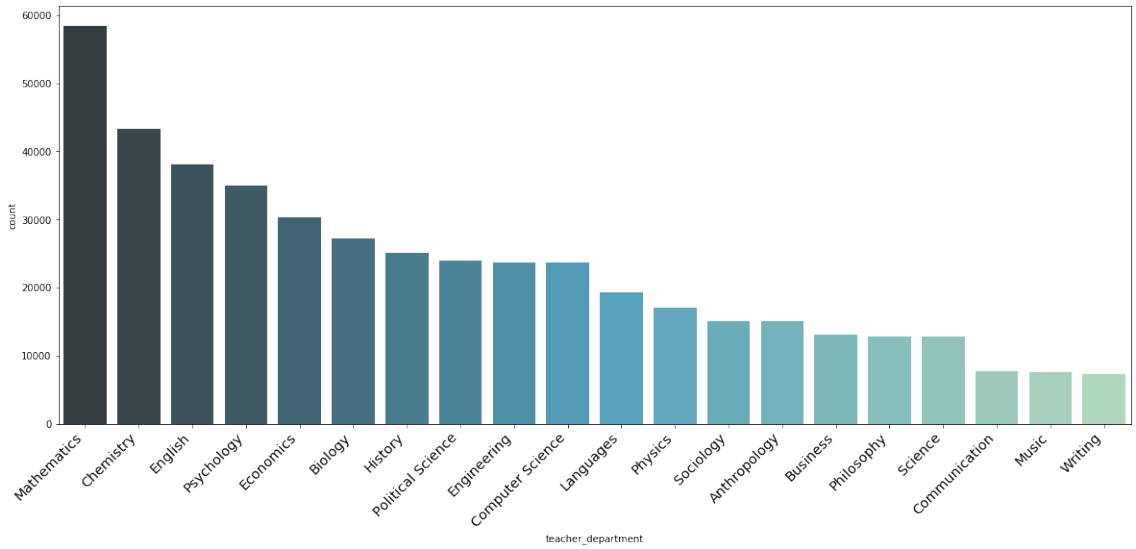

Figure 6: Distribution ratings by department (top 20)

Most used words Since this thesis is mainly based on the use of Natural Language Processing techniques, we also want to deepen our knowledge about the textual content of the dataset. Figure 7 shows the most used words in the corpus of the dataset. The terms that appear larger are the most quoted. Not surprisingly, the most used terms are positive terms or expressions, since, as seen in the distribution graphs in Figure 3, ratings tend to be mainly positive.

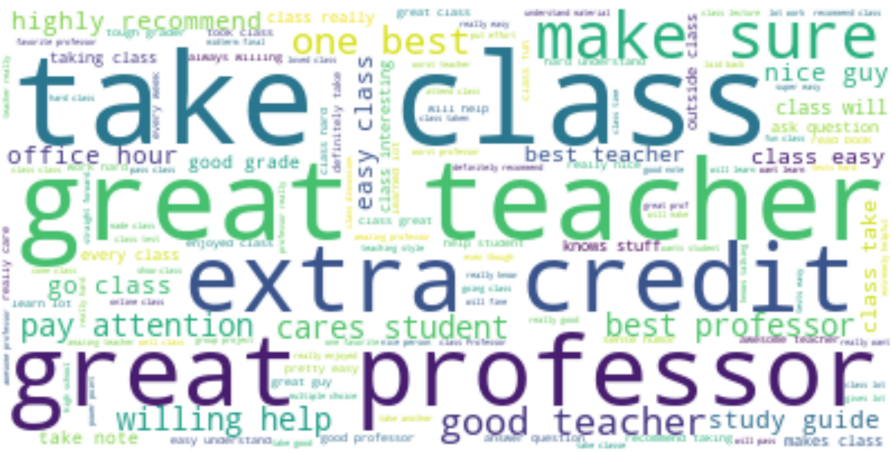

Figure 7: Most used words in corpus

4 Methodology

This section focuses on the processing of data collected in the data harvesting process described in section 3. Initially, an introduction on the methods used for data preparation and preprocessing of textual data is given. An explanation of how the textual data is transformed so that it can be used to answer the various research questions is also given. In addition, an understanding of the method used to select the most representative features to answer each research question is given. Finally, the methodology used to answer the questions presented in the introduction of the thesis is discussed. For each question the results obtained are also shown and discussed.

4.1 Data preparation

Preparing data for analysis is typically a time-consuming process. In fact, according to statistics based on research conducted by CrowdFlower, data scientists spend about 80% of their time preparing and cleaning data for analysis (Press, 2016). However, since our dataset has already been built taking into account its use and since the data comes from a single source, the preparation and cleaning process will not be very demanding.

The first measure of preparation that has been taken is to remove the evaluations that do not report a textual comment since our analysis will be based mainly on NLP. Furthermore, as seen in section 3.2 the data is not balanced in terms of scores. For this reason we will experiment with data balancing techniques.

4.2 Data preprocessing

Text preprocessing is one of the key steps in a text classification problem. In fact, before performing any feature extraction and feature selection operation it is necessary to normalize the text data. The preprocessing stage typically consists of tasks such as conversion to lowercase, stop-words removal, lemmatization, stemming and tokenization. Preprocessing is a very important step in a text classification framework. As shown in previous studies, this step allows to increase the accuracy levels of the models in most cases (Uysal & Gunal, 2014). The removal of stop-words allows to remove terms that most commonly appear in a text and that do not depend on a specific topic (e.g. articles and prepositions). The conversion to lowercase task, on the other hand, allows to remove the difference between words written in lowercase and words written in uppercase, as the meaning does not change, but are still differentiated by a computer. Finally lemmatization allows to transform a word from an inflected form to its canonical form (Lemma; e.g. from successfully, successes, successfulness to success). Since the results of previous studies on the effects of text normalization on the accuracy of classification models have observed that in most cases accuracy is increased, a preprocessing text step has been implemented. The preprocessing stage is divided into five steps. The first step converts text to lowercase, the second step removes numbers, the third step removes punctuation, the fourth step removes stop-words and lastly a lemmatization operation is performed.

4.3 Feature extraction

Since typical algorithms are unable to understand text in its pure form, it is necessary to extract the characteristics of the text and transform them into vectors that the algorithm can interpret. There are several methods to do this. The simplest of all is Bag of Words (BoW). BoW creates a list, called vocabulary, of all the unique words by parsing the entire collection of documents. This allows each document to be represented in vectorial form, where each word is represented with a 1 if the word exists in the document or with a 0 if the term is not present in the document (Trstenjak, Mikac & Donko, 2014, p. 1360). Another representation can be the number of occurrences of a term within a document (Term Frequency; TF) (Trstenjak et al., 2014, p. 1360). However, these two representations do not consider the rarity of terms, because a word that appears less in the text will have lower importance than

frequently recurring terms that often have no value (such as articles and prepositions) (Sewwandi, 2019). A technique that also allows to consider the rarity of terms is TF-IDF (Term Frequency - Inverse Document Frequency). In this technique, TF has the same meaning as in the BoW technique, which can be represented in the following way:

$$TF_{ij} = \frac{n_{ij}}{\mid d_j \mid}$$

where n_{ij} is the number of occurrences of term i in document j, while d_j is the number of terms in document j. Whereas IDF expresses the general importance of the term i in the whole collection of documents. IDF is represented as follows:

$$IDF_i = log \frac{\mid D \mid}{\mid \{d : i \in d\} \mid}$$

where $\mid D \mid$ is the number of documents in the collection, while the denominator is the number of documents containing the term i. TF-IDF can then be represented in the following form:

$$(TF - IDF)_{ij} = TF_{ij} \times IDF_i$$

This technique weighs the frequency of terms present in a document against the prevalence of the term in the collection, making it possible to consider terms that do not appear frequently in the collection.

The TfidfVectorizer processor provided by the Scikit-Learn library was used to extract the features of the dataset corpus. Using the TfidfVectorizer processor the vocabulary of the unique terms of the collection was built. In order to ensure that expressions formed by multiple terms were also considered, a mixture of unigrams, bigrams and trigrams was experimented. In addition, in order to ignore specific terms in the collection (e.g. in our case, terms such as class, professor and student) a maximum document frequency of 0.5 has been set, which means that terms that appear in more than 50% of documents are ignored.

4.4 Feature selection

The resulting vectors from the feature extraction process described in section 4.3 have a great dimensionality due to the fact that (almost) every term in the corpus of the collection is represented as a feature. However, not all terms in the vocabulary generated by the TF-IDF vectorization process are representative and often a large proportion of the vocabulary terms deflect the classification task. Therefore it is

necessary to select a number of features that most represent the original meaning of the document. To do this, the SelectKBest selector provided by the open-source Scikit-Learn library is used. This selector uses the matrix generated by the feature extraction process to select the most representative K features, where K is the number of features to be selected. Chi-square (χ^2), which measures the degree of dependency between two stochastic variables, has been chosen as the scoring system for the selection of the most representative features. χ^2 can be formulated as follows:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the number of observations of type i, E_i are the expected values of type i and n is the number of cells in the table.

However, chi-square does not represent the performance of the model. For this reason, feature percentages from 1% to 100% are tested using a 5 folds cross validation method to find the most representative number of features in each of the tasks described in sections 4.5 and 4.6. Applying this method to each of the tasks allows us to find the most representative features to solve specific problems.

4.5 Comment level processing

The first part of our experiments focuses on the processing of individual evaluations made by students. In this section, we discuss four experiments applied at the comment level. In the first experiment, an attempt is made to distinguish positive comments from negative comments. In the second experiment, besides trying to predict the polarity of the comment, an attempt is made to predict the quality score given by the student. Whereas, similar to the second experiment, in the third experiment, we try to predict the difficulty score given by the student. Finally, in the last experiment, we try to use ML and NLP techniques to distinguish the evaluations of students who have followed the course online or, respectively, physically in the classroom.

4.5.1 Can we distinguish positive from negative comments?

Before we try to predict the quality score given by the student we want to see if it is possible to distinguish positive from negative comments. For this task the quality score is also used, considering evaluations with a score of 1 or 2 as negative evaluations,

while evaluations with a score of 4 or 5 as positive evaluations. Assessments with a score equal to 3 are not considered as the score represents an average or neutral opinion. After removing ratings without a textual comment and ratings with a score equal to 3, 1,453,373 ratings remain, either classified as positive or negative. Figure 8 on page 24 shows the distribution of ratings for each of the two classes. As can be seen in the distribution there is a great disparity between the amount of positive and negative comments. This is due to the fact that, as seen in section 3.2, student assessments tend to be rather upward.

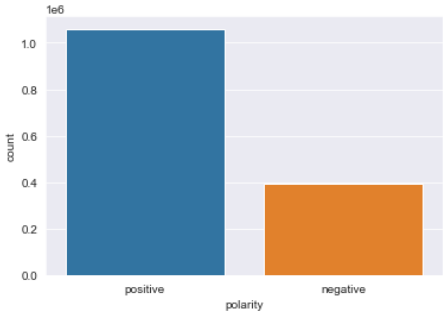

Figure 8: Distribution of evaluations by positive and negative classes

Data imbalance is often a problem in classification tasks, as most algorithms are designed to maximize accuracy and reduce error (Boyle, 2019). When working with imbalanced data, it is very likely that the classification algorithm will prevail over the classes with the highest number of values and ignore the classes with the lowest number of records (Chawla, Japkowicz & Kotcz, 2004). For this reason we decided to evaluate and discuss the differences between training the algorithm with imbalanced data and balanced data. This will expose the problems of using imbalanced datasets.

To determine the training and test datasets we use a random split method, ensuring that the same proportion of classes are present in both datasets. The split ratio between training and test has been set to 70% for training and 30% for testing.

Imbalanced data For the experiment with imbalanced data we took the dataset as it results after the preparation, cleaning, preprocessing and feature extraction processes described in the respective sections 4.1, 4.2 and 4.3. In order to find the most representative features to solve this task with imbalanced data we use the method mentioned in section 4.4. Figure 9 shows the different average scores for the different metric systems obtained by applying the 5 folds cross validation method on the training data. The graph also shows the scores for both in-sample testing and

out-of-sample testing. The best accuracy score is obtained using 10% of the top-K features, meaning about 139,339 features.

Looking at figure 9 we can see a certain instability in the various performances. To better understand what the trend of the different graph curves means, we train the algorithm using 10% of the features with the training dataset and we test it using the test dataset. Table 4 reports the confusion matrix obtained by testing the model using the test data.

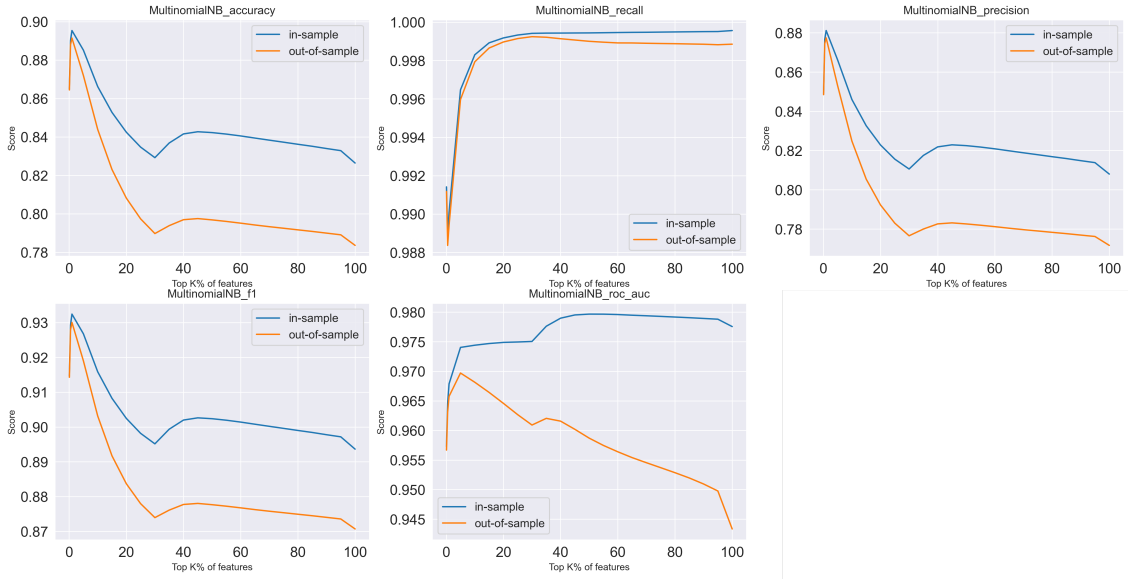

Figure 9: Model scores using different k-values for feature selection for the classification of positive and negative evaluations at comment level using imbalanced data

n = 436,012	Actual: positive	Actual: negative	
Predicted: positive	312,121	37,395	Positive Predictive Value = 0.893
Predicted: negative	5,918	80,578	Negative Predictive Value = 0.932
	True Positive Rate = 0.981	True Negative Rate = 0.683	
	False Negative Rate = 0.018	False Positive Rate = 0.317	

Table 4: Confusion matrix for classification of positive and negative evaluations at comment level using imbalanced data

Looking at the confusion matrix we immediately noticed that the algorithm cannot distinguish between positive and negative comments in an optimal way. In fact the value of False Positive Rate is quite high. As discussed previously, the problem of imbalanced data causes the algorithm to prevail over the class with more

values. When the algorithm trained with imbalanced data will have to predict new data, it will classify it with a very high rate in the class with more values.

Balanced data To balance the two classes we decided to use the down-sampling method. This method reduces the number of values in the class with the most values, bringing the total equal to the class with the least number of values. This results in 393,256 values for each class.

In order to find the most representative features with balanced data we used the same method used for imbalanced data. Figure 10 shows the different average scores for each of the metric systems obtained with the cross validation method. The best accuracy score is obtained using 15% of the features.

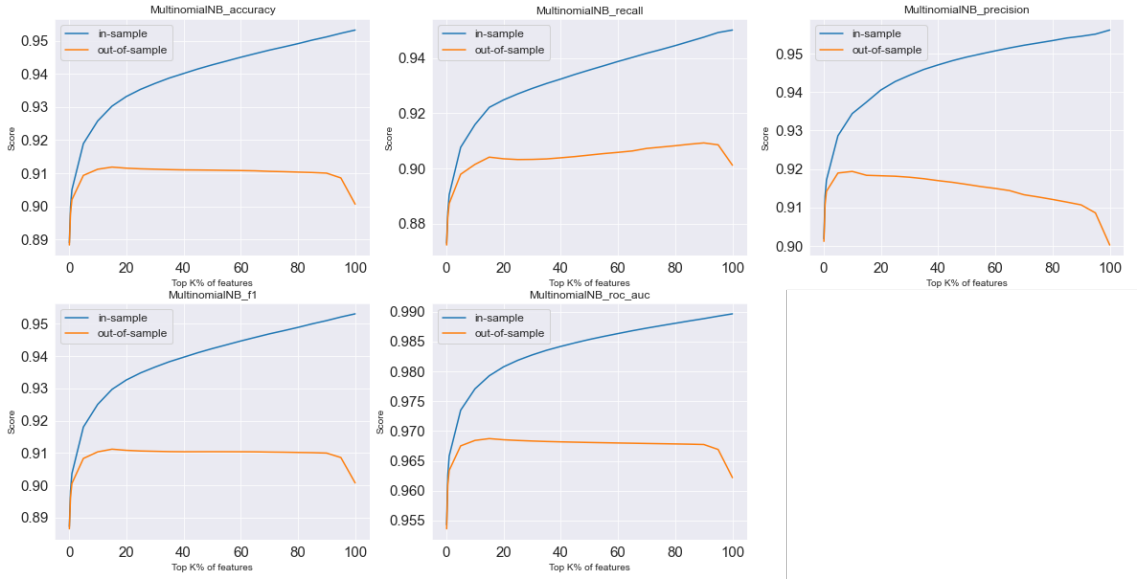

Figure 10: Model scores using different k-values for feature selection for the classification of positive and negative evaluations at comment level using balanced data

The graphs in figure 10 show significantly more stable curves compared to the curves resulting from the use of imbalanced data shown in figure 9. To make a more detailed comparison we train the classification algorithm using 15% of the features with the balanced training dataset and test it with the test dataset. Table 5 reports the confusion matrix for the test data.

n = 235,954	Actual: positive	Actual: negative	
Predicted: positive	105,406	10,813	Positive Predictive Value = 0.907
Predicted: negative	12,546	107,189	Negative Predictive Value = 0.895
	True Positive Rate = 0.894	True Negative Rate = 0.908	
	False Negative Rate = 0.106	False Positive Rate = 0.092	

Table 5: Confusion matrix for classification of positive and negative evaluations at comment level using balanced data

In this case, the confusion matrix reports much more uniform values than the values resulting from the classification with imbalanced data. The classifier still makes classification errors, but of smaller magnitude than those made by the classifier trained with imbalanced data. In fact it is possible to notice a marked difference in the value of False Positive Rate (-0.225Δ).

Therefore it is safe to say that it is possible to distinguish positive comments from negative comments with very high accuracy. It has also been confirmed that the classifier is more effective when trained with balanced data. Table 6 summarizes the top-10 features that represent the most positive and negative evaluations.

Class	Top-10 features
Negative	comment, class, take, hard, test, teacher, dont, doesnt, professor, worst
Positive	class, comment, great, professor, teacher, easy, best, good, really, take

Table 6: Top-10 most representative features for positive and negative evaluations

It can be seen that the classifier clearly distinguishes positive words from negative words such as the term *worst* associated with negative evaluations and the term *great* associated with positive evaluations. However, there are terms such as *professor* and *teacher* that appear among the most associated terms for both classes.

4.5.2 Can we predict the evaluation quality score?

Now that we know that it is possible to distinguish between positive and negative comments with very high accuracy we want to go further and try to predict the clarity score based on the language of the students. This experiment presents a challenge because, unlike a binary distinction (positive or negative), we expect that there are no terms that are specific to a score. Furthermore, reading the RMP assessments carefully, it can be noticed that some assessments are contradictory. In

fact, there are evaluations in which the comment is fully positive, while the given score is very low. For this reason we will try to use the correlations found in section 3.2 to facilitate the experiment.

In contrast to the experiment described in section 4.5.1, evaluations with a score equal to 3 are also considered in this experiment. After removing the evaluations without a textual comment 1,635,693 evaluations remain distributed over the 5 score categories. Figure 11 shows the distribution of the evaluations.

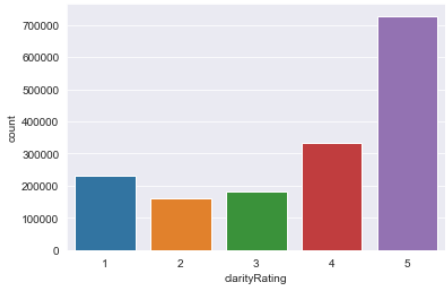

Figure 11: Distribution quality scores

As already seen in section 3.2, the clarity scores are not evenly distributed. In fact there are 727,846 ratings with the clarity score equal to 5 and the remaining 907,847 ratings are distributed over the remaining 4 classes. Hence also in this case the data is highly imbalanced. For this reason we will train the algorithm with both imbalanced and balanced data.

Imbalanced data As in the previous experiment we start to evaluate the performance of the classification algorithm using the imbalanced dataset. Also for this experiment we use the dataset as it results after the preparation, preprocessing and feature extraction phases. We also use the feature selection method described in section 4.4 to find the most representative features to solve this task. Figure 12 shows the different average accuracy scores obtained by applying the cross validation method.

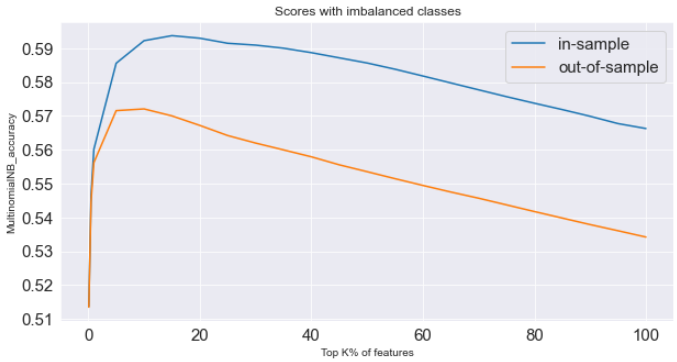

Figure 12: Average accuracy scores using different k-values for feature selection for quality score classification at comment level using imbalanced data

From the graph it can be immediately noticed that the maximum peak of accuracy is relatively high (obtained with 10% of the features), but to better understand what this means we need to investigate further. Table 7 shows the classification report, which shows the precision, recall and F1 values for each of the classes resulting from training the classifier with 70% of the data and testing with the remaining 30%.

	Precision	Recall	F1	Support
1	0.60	0.69	0.64	69,090
2	0.41	0.07	0.13	48,778
3	0.35	0.07	0.11	54,634
4	0.33	0.14	0.19	99,608
5	0.60	0.96	0.74	218,598
Accuracy			0.57	490,708
Macro average	0.46	0.39	0.36	490,708
Weighted average	0.50	0.57	0.48	490,708

Table 7: Report of the classification of quality scores at comment level using imbalanced data

From the classification report we can see that, even if the reported accuracy is relatively high, the classification algorithm does not work precisely. In fact looking at the scores obtained for each class it can be noticed that the algorithm is able to distinguish more or less accurately classes 1 and 5 (which represent the opposite poles), but everything between classes 1 and 5 is distinguished with a lower frequency than it would be by chance. The main reason why this happens is the fact that the data with which the algorithm has been trained are imbalanced, therefore the algorithm prevails the class with the highest number of values. The second reason why the algorithm cannot distinguish between classes 2, 3 and 4 is that, as explained

above, there are no specific terms for each of the classes. For this reason we try to exploit other variables related to the clarity score discovered in section 3.2 to improve the performance of the classifier. Figure 13 shows the average accuracy score obtained taking into account also the features helpfulRating and difficultyRating.

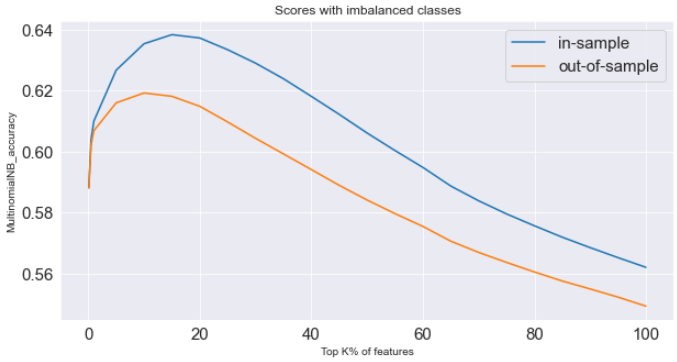

Figure 13: Average accuracy scores using different k-values for feature selection for quality score classification at comment level using imbalanced data and additional features

The maximum accuracy, reached using 10% of the features is about 62% (+5%Δ); to better understand if there are actual improvements we need to investigate further. Table 8 shows the classification report resulting from training the algorithm using 70% of the data and testing with the remaining 30%. The classification report shows the precision, recall and F1 scores for each of the classes.

	Precision	Recall	F1	Support
1	0.70	0.83	0.76	69090
2	0.44	0.26	0.33	48778
3	0.38	0.12	0.19	54634
4	0.39	0.29	0.34	99608
5	0.68	0.89	0.77	218598
Accuracy			0.61	490708
Macro average	0.52	0.48	0.48	490708
Weighted average	0.57	0.61	0.57	490708

Table 8: Report of the classification of quality scores at comment level using imbalanced data and additional features

The recall values of the different classes indicate that there has been a significant improvement in terms of performance. In fact, the proportion of ratings correctly classified in classes 2 and 4 has increased considerably. However, ratings with a score of 3 are still not classified in an acceptable way. Here, too, it is assumed that the

cause is a mix due to the imbalanced data and the neutrality of ratings with a score of 3. We also experimented with other variables correlated with the *clarityRating* feature, such as the features *wouldTakeAgain* and *grade_numerical*, but due to the fact that this data is not available for all evaluations in the dataset, it results in a loss of performance.

Balanced data To determine whether there is a difference between training the algorithm with imbalanced data and balanced data, as in the case of the binary classification described in section 4.5.1, we make an attempt with balanced data. As for the binary classification experiment (positive & negative) we use the down-sampling balancing method. After applying the balancing method there remain 162,495 evaluations for each of the classes.

In order to determine the best accuracy we apply the same feature selection method to find the most representative features. Initially, we try to predict the clarity score using only features extracted from the corpus. Figure 14 shows the average scores obtained by applying the 5 folds cross validation method on top-K features.

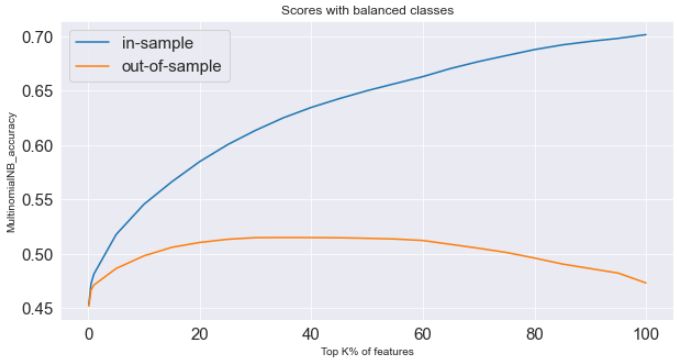

Figure 14: Average accuracy scores using different k-values for feature selection for quality score classification at comment level using balanced data

As can be seen from the graph, the maximum accuracy reached is lower than that achieved using imbalanced data. This is caused by the fact that the accuracy formula only considers the correct classifications. Therefore, if the algorithm has been trained with an imbalanced dataset, it predicts more frequently the most present values. To make a more detailed comparison, table 9 shows the classification report resulting from training with 70% of the data and testing with the remaining 30 per cent.

	Precision	Recall	F1	Support
1	0.55	0.65	0.60	48563
2	0.41	0.37	0.39	48954
3	0.38	0.37	0.38	48837
4	0.41	0.37	0.39	48685
5	0.59	0.61	0.60	48704
Accuracy			0.48	243743
Macro average	0.47	0.48	0.47	243743
Weighted average	0.47	0.48	0.47	243743

Table 9: Report of the classification of quality scores at comment level using balanced data

From the recall scores it can be observed that by training the algorithm with balanced data it is possible to classify the evaluations in their correct score class with a much higher rate than by using imbalanced data. However, the problem that the classifier correctly recognizes the extreme pole ratings (with scores 1 and 5) much better than the ratings with scores 2, 3 and 4 persists. For this reason we try to exploit, in addition to the features extracted from the corpus, the other features correlated with the clarity score mentioned in section 3.2. Figure 15 shows average accuracy scores including, in addition to the TF-IDF matrix, the features helpfulRating and difficultyRating.

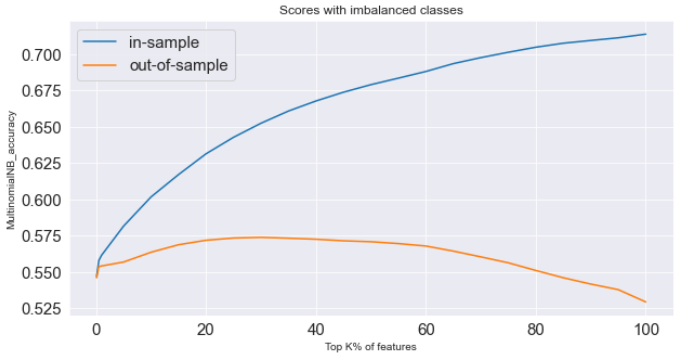

Figure 15: Average accuracy scores using different k-values for feature selection for quality score classification at comment level using balanced data and additional features

The maximum accuracy score achieved ($\sim 57.5\%$) using 30 percent of the most representative features does not show a great increase in performance. To better understand which portion of the evaluations was rated correctly, the classification report resulting from classifier training with 70 percent of the data and testing with

30 percent of the dataset is proposed.

	Precision	Recall	F1	Support
1	0.69	0.73	0.71	48,563
2	0.50	0.43	0.46	48,954
3	0.45	0.42	0.43	48,837
4	0.43	0.44	0.44	48,685
5	0.61	0.69	0.64	48,704
Accuracy			0.54	243,743
Macro average	0.54	0.54	0.54	243,743
Weighted average	0.54	0.54	0.54	243,743

Table 10: Report of the classification of quality scores at comment level using balanced data and additional features

As can be seen from the classification report, the recall scores are much higher than the scores obtained using only features extracted from the imbalanced dataset corpus. The proportion of ratings that have been correctly classified represents almost double the score that would be obtained by randomly assigning a class to the ratings (by chance). It can therefore be inferred that balancing the data, in our case, increases the model's performance. It can also be deduced that the use of features difficultyRating and helpfulRating, in addition to the features extracted from the corpus of the dataset using the TF-IDF vectorisation method, helps to increase the performance of the model, thus confirming the assumptions made in the hypotheses number 2 and 3. Table 11 summarises the top-10 most representative features for each quality rating category.

Quality score	Top-10 features
1	teach, student, doesnt, dont, professor, ever, teacher, worst, take, comment
2	like, grade, teacher, time, dont, lecture, take, test, hard, comment
3	teacher, take, nice, lot, really, easy, hard, good, test, comment
4	take, lot, test, really, professor, easy, teacher, good, great, comment
5	good, awesome, really, take, easy, best, professor, teacher, great, comment

Table 11: Top-10 most representative features for quality score class

Analyzing carefully the table with the top-10 features, it can be observed that in the classes at the extreme poles (1 and 5) there are mainly the same terms that appeared in the classification of positive and negative evaluations. The terms associated with these two classes are mainly very polar. To better visualize the meaning of the table, a simple analysis can be made where the number of positive

terms and the number of negative terms for each rating category are counted. The analysis can be summarized as follows:

- Class 1: 3 negative terms
- Class 2: 2 negative terms
- Class 3: 1 negative term and 3 positive terms
- Class 4: 3 positive terms
- Class 5: 5 positive terms

It follows that the closer the evaluation score get to the poles, the more positive or negative terms there are. The central class (3), being neutral, is represented by both positive and negative terms.

Through this experiment we have collected enough evidence to accept hypothesis 1a. In fact, it has been observed that using purely NLP and ML techniques for the analysis of the students' language it is possible to classify the students' assessments in their correct score classes with a much higher frequency than the statistical randomness frequency. It is also possible to accept the hypotheses number 2 and 3, as it was possible to observe an increase in accuracy of the model by including in the analysis the helpfulness score and the level of difficulty perceived by the students.

4.5.3 Can we distinguish difficult from easy subjects?

As for the prediction of the quality score, we begin in this case too with a binary classification to determine whether it is possible to distinguish difficult from easy subjects. To perform this experiment we use the feature difficultyRating which is transformed into a binary label. Evaluations with a difficulty rating of 1 or 2 are considered to be evaluations that relate to an easy subject. Evaluations with a score of 4 or 5 are considered to be evaluations of a difficult subject. Whereas evaluations with a score of 3 are not considered in this experiment. As seen in section 3.2 the difficulty scores have a normal distribution, so the binary classes have about the same number of assessments each. Also in this experiment the feature selection method is applied using the 5 folds cross validation function in order to find the most representative features to answer the question. Figure 16 shows the performance of the model using different metrics.

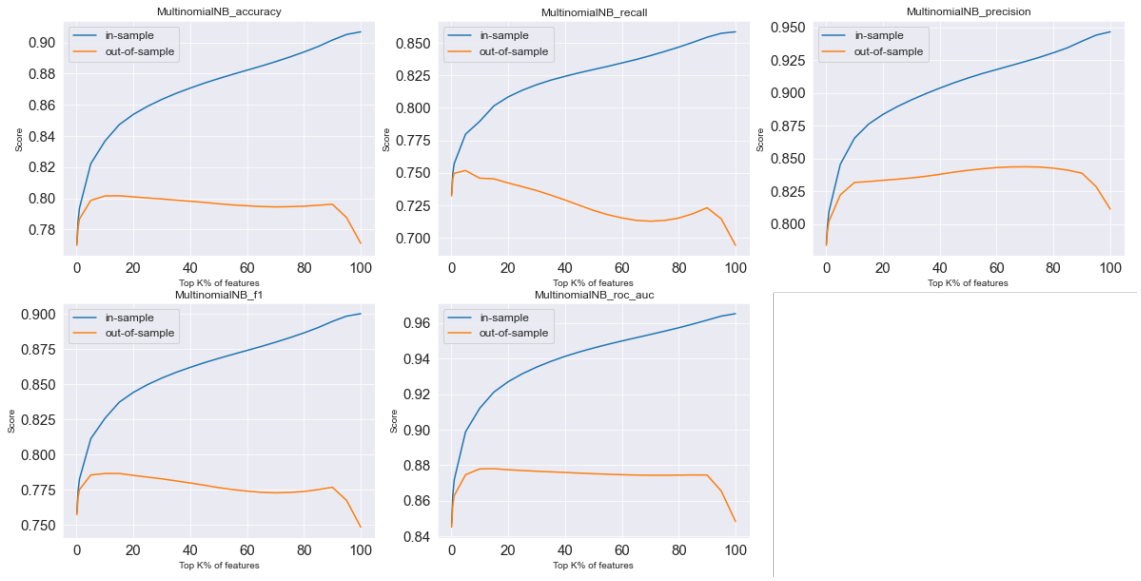

Figure 16: Model scores using different k-values for feature selection for the classification of easy and difficult subjects at comment level

As can be seen from the graphs in figure 16 the maximum accuracy of about 80% is obtained using 15% of the most representative features. To better understand the results of the classification is proposed the table 12 that reports the confusion matrix obtained by testing the algorithm using the test portion of the data.

n = 344,433	Actual: difficult	Actual: easy	
Predicted: difficult	121,719	30,120	Positive Predictive Value = 0.802
Predicted: easy	47,746	144,848	Negative Predictive Value = 0.752
	True Positive Rate = 0.718	True Negative Rate = 0.828	
	False Negative Rate = 0.282	False Positive Rate = 0.172	

Table 12: Confusion matrix for classification of easy and difficult subjects at comment level

From the confusion matrix it can be seen that the model is able to predict more accurately the evaluations concerning an easy subject. In fact, the recall value for the class *easy* (true negative rate) is about 11% higher than the recall value for the class *difficult* (true positive rate). To visualize the words most associated to each class, the weights assigned by the model to the features are used. Table 13 reports the 10 features with the highest weight coefficient for each class.

Class	Top-10 features
easy	easy, great, teacher, professor, good, take, really, best, awesome, helpful
difficult	hard, professor, take, teacher, good, great, students, dont, really, tests

Table 13: Top-10 most representative features for easy and difficult subjects

As can be seen in table 13 the most representative term chosen by the model for the class *easy* is precisely the term *easy*. For the class *easy* the model has also chosen other predominantly positive terms. For the class *difficult* the most representative term chosen by the model is the term *hard*. However, there are still some terms that appear among the features with the highest weight coefficient of both classes (i.e. the terms *good*, *professor* and *teacher*). The results show that it is possible to distinguish between easy and difficult subjects using NLP and ML techniques.

4.5.4 Can we predict the level of difficulty perceived by students?

The hypothesis number 1b assumes that it is possible to predict the difficulty of the subject perceived by the student based on the comment provided in the evaluation. Like the experiment that focuses on the prediction of the clarity score, described in section 4.5.2, this experiment also presents itself as a challenge, as it is expected that there are no specific terms for each scoring class.

Given the much better results obtained using a balanced dataset, we adopt this approach right away. Also to answer the question whether it is possible to predict the level of difficulty of the subject perceived by the student, the feature selection method mentioned in section 4.4 and used in the other tasks is applied. Figure 17 shows the graph that reports the average accuracy scores obtained by applying the 5 folds cross validation method on the top-K features obtained by applying the chi-squared test method.

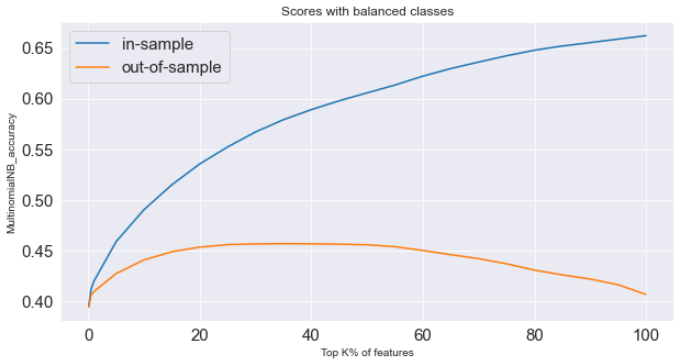

Figure 17: Average accuracy scores using different k-values for feature selection for difficulty score classification at comment level

As can be seen from the graph the maximum accuracy obtained (about 45.5 percent) is not very high. But to better understand how the algorithm classifies the evaluations in the 5 classes of difficulty score it is necessary to investigate further. Table 14 proposes the classification report obtained by testing the classifier using the test dataset composed by 30% of the data.

	Precision	Recall	F1	Support
1	0.48	0.42	0.45	65,887
2	0.34	0.39	0.36	66,035
3	0.31	0.20	0.25	65,930
4	0.34	0.34	0.34	65,923
5	0.51	0.67	0.58	65,832
Accuracy			0.40	329,607
Macro average	0.40	0.40	0.40	329,607
Weighted average	0.40	0.40	0.40	329,607

Table 14: Report of the classification of difficulty scores at comment level

From the classification report it can be immediately understood that the situation is very similar to the quality score prediction described in section 4.5.2. In fact, evaluations with a difficulty score of 1 and 5 are classified in a relatively good way. On the other hand, evaluations with an average difficulty score have considerable resistance to being classified correctly. For this reason we try to use the clarityRating and helpfulRating features in order to find out whether it is possible to improve the performance of the algorithm. So in addition to the features resulting from the feature extraction process we include the features clarityRating and helpfulRating. Figure 18 shows the performance in terms of accuracy.

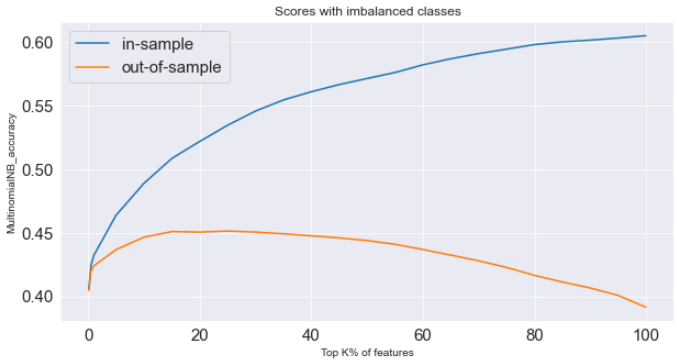

Figure 18: Average accuracy scores using different k-values for feature selection for difficulty score classification at comment level using additional features

As can be seen from the graph there is very little difference in terms of accuracy compared to the graph generated by using only the features extracted from the corpus. Apparently the features clarityRating and helpfulRating have a very small effect on the model accuracy in this case. But to better understand if there are any other positive effects produced by the use of additional features it is necessary to compare the performance in more detail. Table 15 shows the classification report. Also in this case the classification report results from training with 70% of the data and testing with the remaining 30% of the dataset.

	Precision	Recall	F1	Support
1	0.43	0.56	0.48	65,887
2	0.34	0.36	0.35	66,035
3	0.31	0.28	0.29	65,930
4	0.37	0.23	0.29	65,923
5	0.58	0.63	0.60	65,832

Accuracy			0.41	329,607
Macro average	0.40	0.41	0.40	329,607
Weighted average	0.40	0.41	0.40	329,607

Table 15: Report of the classification of difficulty scores at comment level using additional features

A detailed analysis of the classification report shows that the proportion of ratings categorized correctly in the correct difficulty score class has increased slightly, but there is still no significant increase. So we deduce that it is easier, though still not very accurate, to predict the quality score in comparison to the difficulty score.

To better understand what may be the source of this difficulty, the most associated terms for each of the scoring classes are analyzed. Table 16 shows the top-10 most

informative features for each scoring class.

Difficulty score	Top-10 features
1	test, good, really, best, take, professor, great, teacher, comment, easy
2	make, test, take, really, good, professor, easy, teacher, great, comment
3	lecture, make, student, take, really, good, teacher, professor, great, comment
4	student, lot, test, good, take, great, teacher, hard, professor, comment
5	student, ever, worst, dont, professor, test, teacher, take, hard, comment

Table 16: Top-10 most informative features for each difficulty score class

As can be seen in the table containing the top-10 most representative features for each of the classes, classes 1 and 2 contain terms that refer to the ease of the subject taught by the teacher such as the term *easy*. Classes 3 and 4 are also associated, along with other terms, with the term *hard* which refers to the difficulty of the subject taught by the teacher. However, for class 3 the algorithm does not use terms referring to the difficulty/easiness of the subject.

On the basis of the results obtained, it is possible to evaluate hypothesis 1b. The recall scores obtained show that the algorithm is able to recognize relatively well evaluations with a high or low difficulty score. However, evaluations with a difficulty score [2, 4] are not classified with the same accuracy. However, it can be observed that the portion of correctly identified positive identifications (precision) for each class is higher than the value of statistical randomness. Therefore it is possible to accept hypothesis 1b.

4.5.5 Can we distinguish online classes?

RateMyProfessors also allows students who have taken an online version of a course to evaluate the instructor. In this experiment we want to try to use NLP and ML techniques to distinguish the assessments of students who have taken an online course or in class.

Since on RateMyProfessors, when creating a new assessment, it is possible to specify whether the course to be assessed was taken online, the target data is already present in our dataset. Figure 19 shows the distribution of assessments according to their class (online & not online).

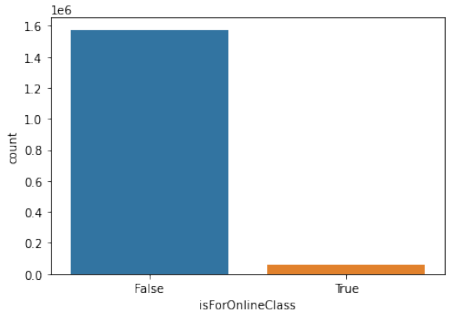

Figure 19: Online class distribution

In our dataset, ratings for courses taken online are clearly scarce in quantity, while ratings for courses taken in class are in great abundance. However, a short analysis has shown that the isForOnlineClass flag is not always used correctly in the ratings. In fact, several evaluations were found whose content of the comment reported that the student followed the online course, while the isForOnlineClass flag was not positive. For this reason we have defined buzzwords such as *online*, *Coursera* and *YouTube* which if they appear in the comment, most likely indicate that the course was followed online.

By applying this method, it was possible to increase the portion of ratings for online courses by about 60,000 units. However, the dataset still shows a great imbalance between classes. Given the much better results obtained using a balanced dataset it was decided to apply the down-sampling method to balance the classes. After balancing the classes there remain 125,287 evaluations for each class, of which 70% are used for training and the remaining 30% for testing.

In order to find the most representative features to answer the question whether it is possible to distinguish the ratings for a course held online we apply the feature selection method with the 5 folds cross validation function. Figure 20 shows the performance of the model in terms of accuracy, recall and precision for the top-K% features.

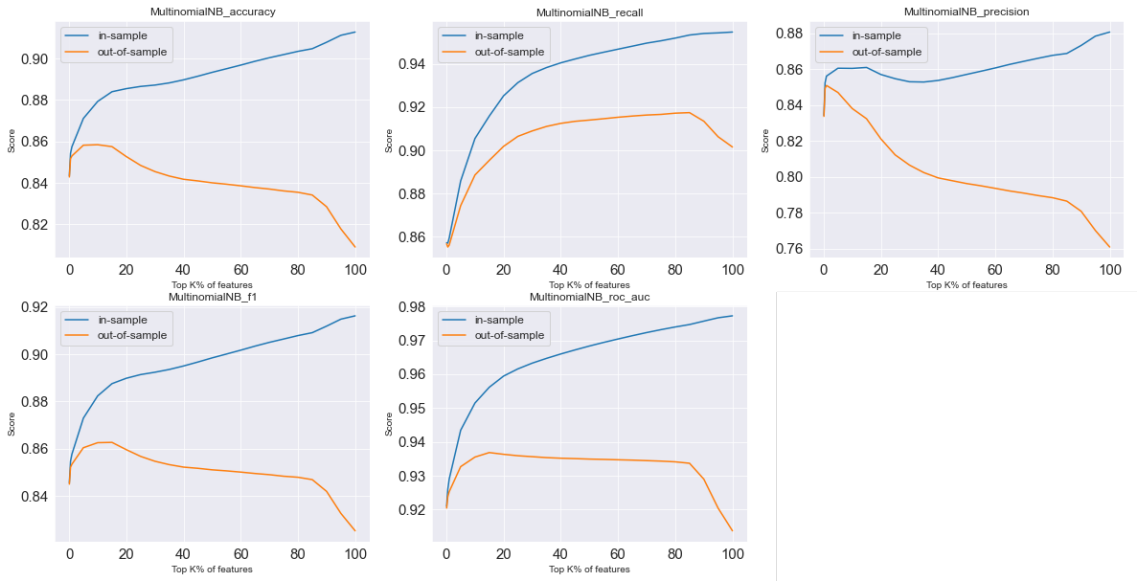

Figure 20: Model scores using different k-values for feature selection for the classification of online and not-online classes

The maximum accuracy score (86 percent) is achieved using 10 percent of the top-K features. Table 17 shows in detail the confusion matrix that reports the results of the classification. The confusion matrix clearly shows that the classification algorithm is able to distinguish the ratings for courses taken online. The false positive and false negative rates are relatively high, however, the evidence shows that the distinction between online and offline courses is significant.

n = 75,173	Actual: online	Actual: not online	
Predicted: online	32,857	7,967	Positive Predictive Value = 0.805
Predicted: not online	4,710	29,639	Negative Predictive Value = 0.863
	True Positive Rate = 0.875	True Negative Rate = 0.788	
	False Negative Rate = 0.125	False Positive Rate = 0.212	

Table 17: Confusion matrix for classification of online and not-online classes

To better understand at a concrete level the features used for this task, we propose table 18 that summarises the most representative features for each of the classes. The features are listed in order of importance.

Class	Top-10 features
not online	helpful, teacher, great, good, take, really, easy, students, best, hard
online	online, easy, take, great, work, took, assignment, tests, course, teacher

Table 18: Top-10 most representative features for online and offline classes

The most significant term for assessments concerning courses held in class is the term *helpful* which could mean the helpfulness of the teacher towards the students during and after class, which is not always possible for online courses. By contrast, it is not surprising that the most important term for assessments regarding online courses is precisely the term *online*.

To get consistent results, the results reported above are compared with the results of the trained model using the feature *isForOnlineClass* already present in the dataset and without considering the feature extracted using buzzwords. Figure 21 shows the results of the feature selection and cross validation method for the case of using the feature extracted from buzzwords and for the case of using only the pre-existing feature *isForOnlineClass*.

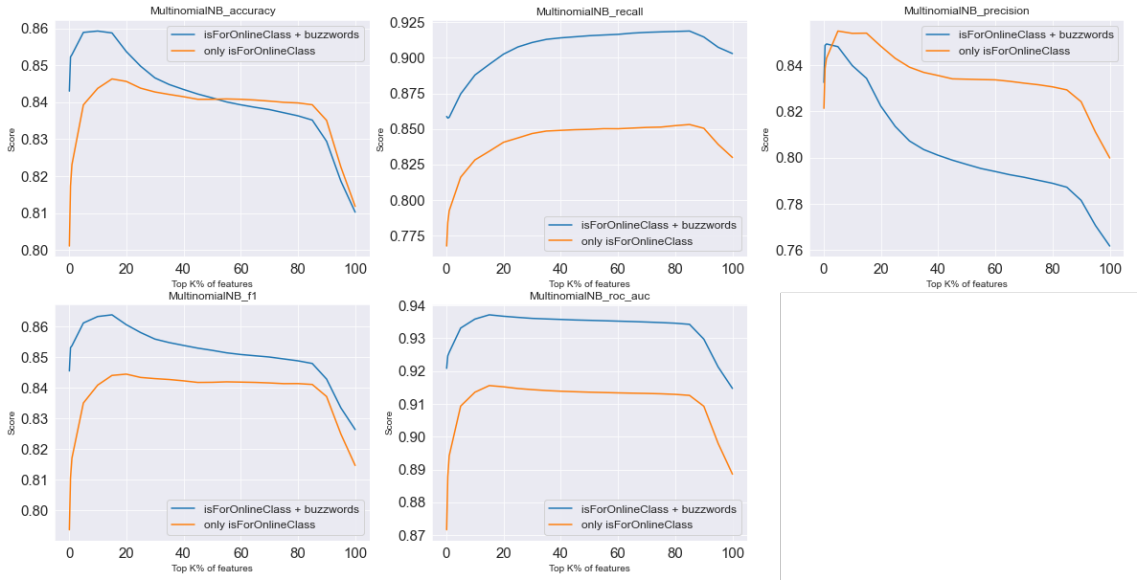

Figure 21: Model scores using different k-values for feature selection for the classification of online and not-online classes using only the preexisting target feature

The graph lines shown in figure 21 primarily show a worse performance when only the pre-existing *isForOnlineClass* feature is used. However, to better understand how evaluations are classified into the classes *online* and *not online*, it is necessary to analyze the confusion matrix shown in Table 19.

n = 36,539	Actual: online	Actual: not online	
Predicted: online	14,816	3,180	Positive Predictive Value = 0.823
Predicted: not online	3,594	14,949	Negative Predictive Value = 0.806
	True Positive Rate = 0.805	True Negative Rate = 0.825	
	False Negative Rate = 0.195	False Positive Rate = 0.175	

Table 19: Confusion matrix for classification of online and not-online classes using only the preexisting target feature

As can be seen from the confusion matrix the value of true positive rate is about 7% lower than in the first case, while the value of true negative rate is slightly higher in the second case. This means that by using the target extraction method based on the use of buzzwords it is possible to better recognize the evaluations that concern an online course. The general decrease in performance when only the pre-existing feature *isForOnlineClass* is used can be mainly due to the fact that, as discussed before, there are many ratings whose comment content concerns an online course, while the *isForOnlineClass* feature is negative.

Therefore, the answer to the question whether it is possible to distinguish between assessments concerning an online course is yes. In fact, the results show an accuracy of about 86%, which is significant evidence. Thus, it is possible to accept the hypothesis number 5.

4.6 Professor level processing

The second part of our experiments focuses on teachers. In this section we discuss 2 experiments applied at teacher level. In the first experiment, similar to what is discussed in section 4.5.1, we try to distinguish good teachers from bad teachers, while in the second experiment we try to estimate the teacher's teaching quality perceived by students. For these two experiments we use the dataset containing information about the teachers of the different schools. In order to be able to use the comments of the student assessments, a feature called *all_comments*, which contains all the textual comments of the assessments related to the teachers is added to the teachers dataset.

4.6.1 Can we distinguish good from bad teachers?

Similar to the experiment described in section 4.5.1, in this experiment we try to distinguish good professors from bad professors. For this task the feature *averageratingscore* is transformed into a binary label, where teachers with an average score below 3 are considered bad teachers, while teachers with an average score above 3 are considered good teachers. After removing the teachers who do not have any ratings, 134,221 teachers remain distributed in the classes good and bad, 70 percent of which will be used for model training and the remaining 30 percent will be used for testing. Figure 22 shows the distribution of teachers in the classes good and bad.

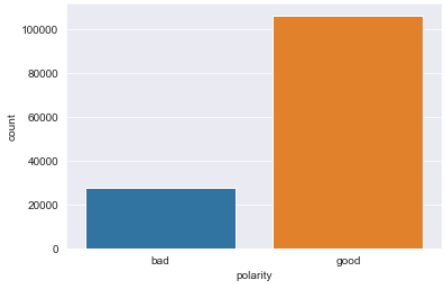

Figure 22: Distribution of good and bad professors

In this case, as in the case of the classification of positive and negative evaluations, good and bad professors are not balanced. Given the clearly better results obtained using a balanced dataset in previous experiments, a balanced dataset is used directly to train the algorithm. The down-sampling method is also used to balance the dataset. After balancing the dataset 27,843 teachers remain for each class.

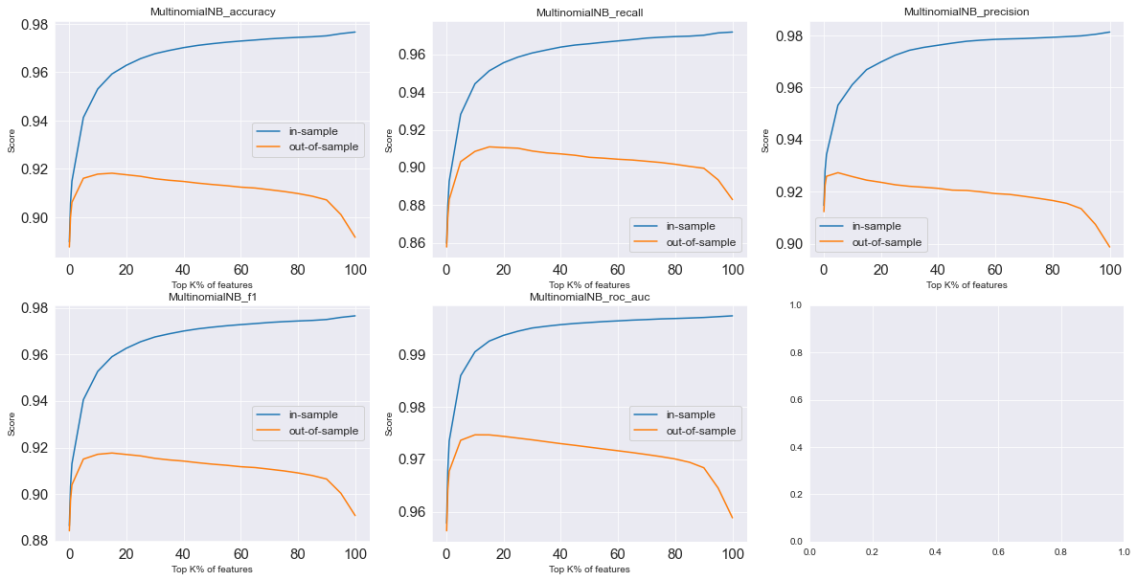

Figure 23: Model scores using different k-values for feature selection for the classification of good and bad professors

Figure 23 shows the performance of the model by applying the 5 folds cross validation method in order to find the most representative features. The maximum accuracy of about 92% is reached by using 15% of the most representative features. Table 20 shows the confusion matrix obtained from algorithm training using 70% of the dataset and 15% of the most representative features and from testing with the test dataset.

n = 16,706	Actual: good	Actual: bad	
Predicted: good	7,330	854	Positive Predictive Value = 0.899
Predicted: bad	1,021	7,501	Negative Predictive Value = 0.880
	True Positive Rate = 0.878	True Negative Rate = 0.899	
	False Negative Rate = 0.122	False Positive Rate = 0.101	

Table 20: Confusion matrix for classification of good and bad professors

As can be seen in the confusion matrix in Table 20 the recall value for the class good (true positive rate; 87.8%) and the recall value for the class bad (true negative rate; 89.9%) are very close.This means that the algorithm can clearly distinguish good professors from bad professors.

Table 21 shows the most representative terms for each class. The most representative terms are obtained by consulting the weight coefficient for each feature.

Class	Top-10 features
bad	hard, worst, dont, doesnt, time, like, questions, work, know, never
good	best, helpful, lot, work, help, one, interesting, nice, always, fun

Table 21: Top-10 most representative features for good and bad teachers

As can be seen from the most representative terms, for the class *bad* are listed mainly negative terms such as *worst* and *hard*. In contrast, for the class *good* the terms listed are mainly positive terms such as *best* and *nice*. Interestingly, unlike the classifications discussed in previous experiments, in this case the most representative terms are much more distinct for each of the classes. In fact in this case there is only the term *work* which is repeated in both classes.

The answer to the question whether it is possible to distinguish good professors from bad professors using NLP and ML techniques is therefore yes.

4.6.2 Can we predict the professor's overall score?

Similar to the experiment discussed in section 4.5.2, in this experiment we try to predict the quality of teaching perceived by students at the teacher level using NLP and ML techniques. In the teacher dataset there is the feature *averageratingscore* that represents the average of all the quality and helpfulness scores reported in the teacher's assessments. This feature takes the form of a real number between 1 and 5. In order to answer the question of this experiment using a classification algorithm it is necessary to transform the teacher's average score into a categorical target. For this reason it was decided to divide teachers into four categories, where teachers with an average score between 1 and 2 belong to category 1, teachers with an average score between 2 and 3 belong to category 2 and so on up to category 4. Figure 24 shows the distribution of teachers within the 4 categories.

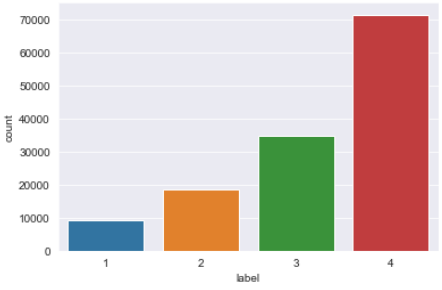

Figure 24: Distribution of professors into 4 quality score classes

Again, as in previous experiments, we are confronted with imbalanced data. Given the considerably better results obtained in the previous experiments with balanced data, in this experiment we adopt the down-sampling method as well. After applying the down-sampling method, 9,191 professors remain for each scoring class.

In order to find the most representative features to perform this experiment, as in previous experiments, we use the feature selection method based on the chi-squared test. Figure 25 shows the average accuracy scores obtained by applying cross validation on the top-K% features resulting from the feature selection method.

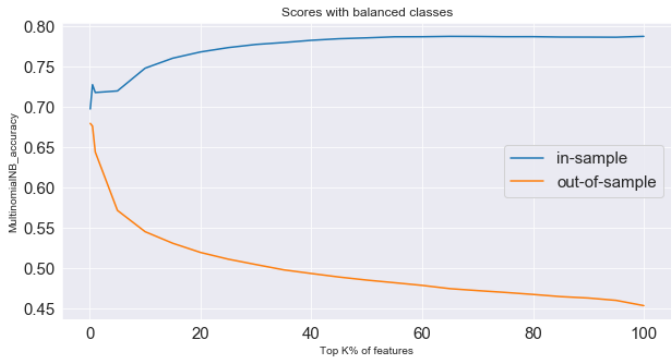

Figure 25: Average accuracy scores using different k-values for feature selection for overall score classification at professor level

As can be seen from the graph in figure 25 the maximum accuracy of about 68% is reached using 0.1% of the features, which translates into about 1,720 features. To better understand how teachers are classified in each class, it is necessary to analyze the results further. Table 22 shows the classification report resulting from model training using 70% of the data and testing with the remaining 30%.

	Precision	Recall	F1	Support
1	0.77	0.74	0.76	2,823
2	0.56	0.62	0.59	2,695
3	0.60	0.59	0.60	2,758
4	0.81	0.77	0.79	2,754

Accuracy			0.68	11,030
Macro average	0.69	0.68	0.68	11,030
Weighted average	0.69	0.68	0.68	11,030

Table 22: Report of the classification of overall score at professor level

As shown in the classification report, the correct classification rate for each class is relatively high. It can also be observed that, as in the experiment discussed in section 4.5.2, the scoring classes at the poles report significantly higher recall and

precision values than the middle classes. The cause of this behavior can be traced back to the fact that, as can be seen in table 23, for the classes representing the lowest and the highest score (1 and 4) there are clearly positive and clearly negative terms such as *worst* and *hard* for the lowest scoring class.

Overall score	Top-10 features
1	worst, dont, doesnt, hard, never, ever, even, know, time, teach
2	lectures, hard, questions, dont, work, like, time, tests, doesnt, easy
3	lectures, dont, help, nice, tests, lot, work, hard, easy, great
4	best, helpful, lot, interesting, work, one, fun, help, easy, great

Table 23: Top-10 most informative features for each overall scoring class

In addition to the features extracted from the corpus of the dataset, we now want to include the gender of the teacher in order to evaluate what is assumed in hypothesis number 4. Indeed, hypothesis number 4 assumes that by including the gender of the teacher in the process it is possible to increase the performance of the model. To do this we use the feature *professor_gender* extracted by us using the first name of the teacher and described in section 3.2. In order to correctly test the hypothesis it is necessary to remove from the dataset the professors whose gender extracted from the first name is unknown or androgynous. This results in a reduction in the number of teachers for each class by 1,614 units.

Figure 26 illustrates the average scores obtained from the cross validation applied to the feature selection method including the *professor_gender* feature. From the graph no difference can be observed between using only the TF-IDF matrix and adding the *professor_gender* feature. Therefore, it is necessary to analyze in more detail the classification report resulting from the separate training and testing.

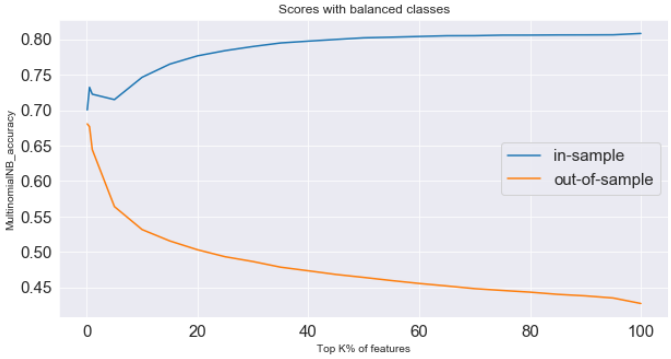

Figure 26: Average accuracy scores using different k-values for feature selection for overall score classification at professor level including feature *professor_gender*

Table 24 shows the classification report resulting from the training with 70% of the data and the testing with the remaining 30%. As can be seen, the accuracy score obtained is exactly the same in both cases. However, comparing the precision and recall scores of both cases, a slight difference can be observed. Generally it can be noticed that by including the feature *professor_gender* there is a deterioration in performance. However, the drop in performance is not necessarily caused to the use of the feature *professor_gender*, but more likely it is caused by the loss of data due to the selection of professors whose gender derived from the first name is not unknown or androgynous. So in this case, the use of the gender of the teacher does not seem to have any effect on the performance of the model.

	Precision	Recall	F1	Support
1	0.74	0.79	0.76	2,262
2	0.64	0.54	0.58	2,327
3	0.56	0.65	0.60	2,214
4	0.81	0.75	0.78	2,290
Accuracy			0.68	9,093
Macro average	0.68	0.68	0.68	9,093
Weighted average	0.68	0.68	0.68	9,093

Table 24: Report of the classification of overall score at professor level including feature *professor_gender*

By including the gender of the professor in the analysis it was not possible to observe a significant difference in the results. In fact, we suppose that the differences shown in the results are caused by the reduction of data due to the selection of professors whose gender is not unknown or androgynous. Therefore hypothesis number 4 cannot be accepted.

4.7 Topic detection

The results presented in the previous sections have shown that it is possible to use NLP and ML techniques to accurately classify positive and negative evaluations, evaluations of online courses as well as good and bad professors based on student language. Although these are already interesting results, we are also interested in giving a little more context to the results obtained by the various classifications. For this reason we want to detect which topics are most discussed in the evaluations of a course held online. In fact we want to find out if there are any main topics that

are discussed when students evaluate an online course, such as topics about online communication or topics about digital material management.

In this experiment we are therefore focusing specifically on evaluations concerning online courses.

Since we do not have a list of topics to look for in the online course evaluations, it is necessary to treat this experiment as an information extraction task. Therefore it is not possible to rely on the typical machine learning training and testing process. Instead, this experiment needs an approach based on clustering and therefore unsupervised learning, where the algorithm uses the same information contained in the documents to learn (Wartena & Brussee, 2008). This task consists mainly of two steps. The first step consists in extracting the terms used in the corpus (building the dictionary), while the second step consists in identifying clusters (topics) formed from the dictionary terms (Wartena & Brussee, 2008).

The dictionary is built first by dividing the documents into sentences, then tokenizing the sentences in terms. Each token is lemmatized and tagged with the corresponding part of speech tag (POS tagging) in order to build meaningful bigrams and trigrams. This allows for example to construct the "give extra credit" trigram where the POS tags of the individual tokens are considered. If the POS tags of the tokens are not taken into account, bigrams or trigrams such as "school highly" could happen, which may be of great importance for the algorithm, but would not make great logical sense. Since we are also interested in semantics for topic detection, it is important that bigrams and trigrams are constructed correctly.

To identify clusters of terms in order to discover the main topics we use the generative statistical model Latent Dirichlet Allocation (LDA). LDA allows to consider each single document as a set of topics and to understand the semantic meaning of the text by analyzing the similarity between the distribution of terms in a document and that of a specific topic (Blei, Ng & Jordan, 2003). LDA assumes that there are a defined number of topics in the entire collection. Each topic is defined as a set of terms from the dictionary extracted from the document collection (Blei, 2012, p. 78).

To implement topic detection using LDA we use the LDA model provided by the Gensim[10] library for Python.

In order to identify topics it is necessary to define the number of topics (k)

[10] Gensim

to be identified. A low k value allows to identify more general topics, but also presents the risk of generalizing the topics too much and therefore not being able to distinguish specific topics. A high k value allows to identify more specific topics, but this, depending on the data used, also carries the risk that the topics identified are too similar to each other and therefore not easily distinguishable. Therefore it is necessary to find the optimal number of topics to identify.

To find the optimal value of k we build multiple LDA models with different values for the number of topics. To understand which model clusters the topics optimally, we compare the coherence score of each model. Figure 27 shows the coherence score for different k-values.

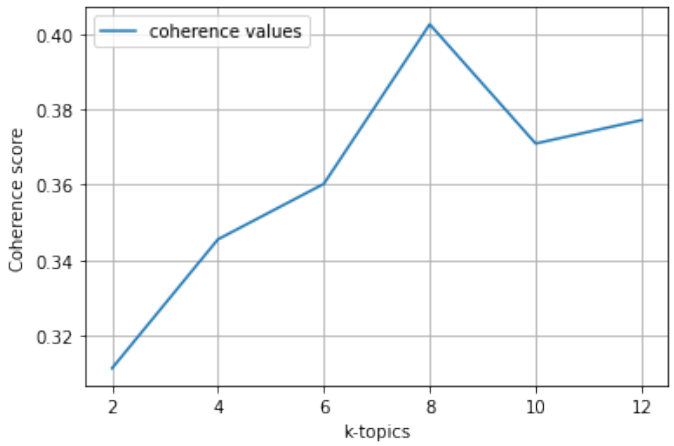

Figure 27: Coherence score for different k-values

As can be seen from the graph in figure 27 the highest coherence score is obtained with $k = 8$, so we build the final LDA model using $k = 8$. The graph in figure 28 shows the topics identified by the LDA model. Each bubble represents a topic. The size of the bubble represents the prevalence of the topic within the collection. A good model should show well distributed bubbles in the four quadrants and there should be no overlying bubbles, just as shown in figure 28.

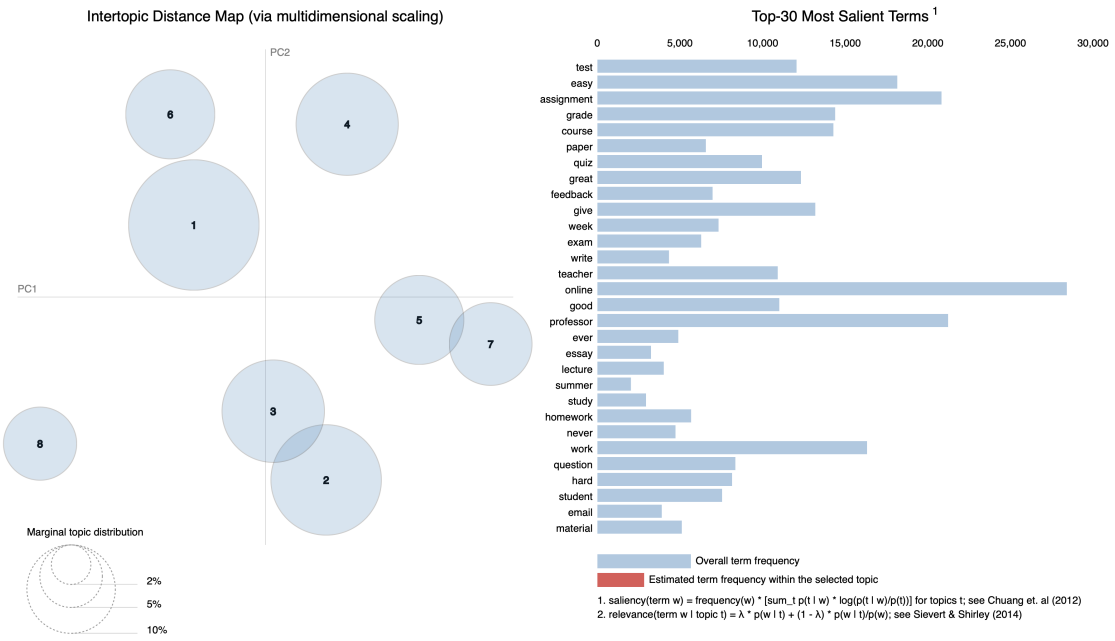

Figure 28: Topic clustering map and top-30 most salient terms

To the right of the clusters graph in figure 28 are also shown the most salient terms in the collection, but to better visualize the terms of each topic identified is proposed table 25 showing the 10 terms that most contribute to the formation of the topic.

As discussed previously, the Latent Dirichlet Allocation, in simple words, groups together the terms most conducive to the formation of a topic, therefore the results produced may not appear clear and require human interpretation. Thus we attempted to give a context to the terms of each identified topic by deriving an assumption of the topic from the terms. In table 25 it is possible to read our assumptions about the topics.

Topic number and topic assumption	Weight	Top-10 terms	Topic number and topic assumption	Weight	Top-10 terms
	5.2%	online		7.7%	assignment
	4.5%	professor		3.6%	week
	2.4%	work		3.0%	online
	2.4%	course		3.0%	grade
1	2.3%	great	5	2.8%	work
	2.0%	teacher		2.5%	time
good and helpful teacher	2.0%	helpful	weekly assignments	2.2%	course
	2.0%	give_help		1.6%	due
	1.8%	good		1.5%	day
	1.6%	lot		1.4%	post
	5.6%	quiz		3.4%	professor
	5.3%	weekly		3.2%	teacher
	3.5%	online		2.6%	work
	2.0%	test		2.6%	hard
2	1.8%	easy	6	2.5%	online
	1.8%	question		2.4%	ever
weekly online quiz	1.6%	textbook	demanding course	2.3%	much
	1.5%	chapter		2.2%	learn
	1.2%	every_week		2.0%	course
	1.1%	require		1.9%	teach
	5.4%	exam		3.0%	feedback
	4.4%	online		2.5%	respond_email_quickly
	3.8%	study		2.2%	grade
	3.5%	question	7	2.0%	paper
3	2.3%	hard		1.9%	write
	2.3%	give		1.8%	encouraging
online final exam	1.9%	time	good and fast feedback	1.6%	creative
	1.7%	extra_credit		1.5%	manageable
	1.7%	pass		1.4%	improve
	1.7%	final		1.4%	ability
	6.5%	never		5.1%	summer
	5.3%	give		2.9%	course
	3.5%	email		2.7%	online
4	2.9%	feedback	8	2.0%	lecture
	2.7%	question		1.9%	material
	2.1%	ask		1.9%	well
poor feedback	1.8%	grade	well organised summer course	1.8%	video
	1.8%	assignment		1.7%	provide_content
	1.7%	thing		1.7%	organize
	1.7%	point		1.7%	textbook

Table 25: Identified topics and terms for each topic

The results of the analysis based on the LDA model show that it is possible to recognize specific topics on faculty performance within the comments of the assessments concerning online courses. Therefore, it can be inferred that the evaluations are not necessarily influenced by the halo effect as assumed in previous studies (Felton et al., 2003; Otto et al., 2008). Consequently it is possible to accept hypothesis number 6.

5 Conclusions

This thesis investigated the use of natural language processing and machine learning techniques for the processing of text data, with the aim of discovering language patterns used by students in the evaluation of their teachers on online platforms such as RateMyProfessors. The main objective of the thesis was to use possible language patterns to answer questions such as whether it is possible to predict the quality of teaching perceived by students based solely on the students' language.

The thesis discussed the various aspects of classic faculty performance assessments conducted internally by schools as well as the aspects and possible bias of online evaluations. In addition, the current state of research in the field of natural language processing was researched, providing also some historical context of this research field. Subsequently, 6 hypotheses were developed based on previous studies and research to be discussed and evaluated on the basis of the results obtained from the various experiments conducted.

In addition, the techniques of data extraction from websites, with which the necessary datasets for the experiments were created, were exposed and discussed. Using these techniques, three datasets containing 1,637,435 evaluations of 134,375 teachers from 605 schools were constructed using a random approach. The various legal aspects involved in the extraction of data from third party websites were also discussed. We also proposed a preliminary analysis of the content of the three datasets to provide a preview of the distribution of the data and the various correlations between the collected variables.

In the part of the thesis where the methodology is discussed the various approaches adopted for the data preparation, the pre-processing of the textual data, the extraction of features from the corpus and the feature selection method based on the chi-squared statistical method for the selection of the most representative features were discussed.

The experiments were divided into two main parts, where in the first part the experiments were focused on the commentary level, while the experiments in the second part focused on the processing of the evaluations at the professor level.

At the commentary level it was shown that it is possible to distinguish positive evaluations from negative evaluations with an accuracy of over 90% based exclusively on the language used by the students in the comments of the evaluations. In addition, it was demonstrated that it is possible to predict the quality of teaching perceived

by students with a baseline accuracy of about 55%. It was also demonstrated that it is possible to distinguish difficult subjects from easy subjects with an accuracy of about 80%. Using this as a starting point, it was also proven that it is possible to predict the level of difficulty perceived by students with a baseline accuracy of about 41%. Finally, for the experiments at the commentary level it was confirmed that it is possible to distinguish between online courses and classroom courses with an accuracy of 86%.

For data processing at the teacher level it was also shown that it is possible to distinguish between good and bad teachers with an accuracy of over 91%.

For each of the experiments, both at commentary and teacher level, the terms most associated with each class were presented to highlight the language patterns. In addition, the difference between using imbalanced and unbalanced data was discussed in the experiments.

Finally, we asked the question whether it is possible to distinguish specific topics concerning the performance of the faculty within the evaluations of courses held online. Using the statistical model latent Dirichlet allocation it was possible to find out that in the evaluations concerning online courses, topics concerning students' learning and the quality of teaching methods are discussed.

Therefore, schools can use publicly available data about their teachers along with natural language processing and machine learning techniques to generate useful information about faculty performance and student learning. However, it is assumed that surveys conducted internally by schools are still an important activity since they remove the possible biases created from online assessments. Therefore, the use of natural language processing and machine learning techniques applied to data from online platforms is to be seen as an extension of internal surveys, which gives the possibility to capture aspects of the performance of the institution and teaching methods that may not be fully considered in internal assessments.

An interesting possible development of this work could be the comparison of the results obtained in the thesis with the results that would be obtained from the processing of data collected internally by schools. It would be interesting to analyze possible discrepancies in the way teachers are evaluated between online evaluations and official evaluations, in order to understand if there are aspects that are discussed differently.

References

Azab, M., Mihalcea, R. & Abernethy, J. (2016). Analysing RateMyProfessors evaluations across institutions, disciplines, and cultures: The tell-tale signs of a good professor. In E. Spiro & Y.-Y. Ahn (Eds.), *Social informatics* (Vol. 10046, pp. 438–453). doi:10.1007/978-3-319-47880-7_27

Blei, D. (2012, April). Probabilistic topic models. *Communications of the ACM*, *55*(4), 77–84. doi:10.1145/2133806.2133826

Blei, D., Ng, A. & Jordan, M. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993–1022.

Boyle, T. (2019, February 4). Methods for dealing with imbalanced data [Medium]. Retrieved April 28, 2020, from https://towardsdatascience.com/methods-for-dealing-with-imbalanced-data-5b761be45a18

Brown, C. L. & Kosovich, S. M. (2015, August). The impact of professor reputation and section attributes on student course selection. *Research in Higher Education*, *56*(5), 496–509. doi:10.1007/s11162-014-9356-5

Brown, M. J., Baillie, M. & Fraser, S. (2009, April). Rating ratemyprofessors.com: A comparison of online and official student evaluations of teaching. *College Teaching*, *57*(2), 89–92. doi:10.3200/CTCH.57.2.89-92

Chawla, N. V., Japkowicz, N. & Kotcz, A. (2004, June). Editorial: Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1–6. doi:10.1145/1007730.1007733

Chowdhury, G. G. (2005, January 31). Natural language processing. *Annual Review of Information Science and Technology*, *37*(1), 51–89. doi:10.1002/aris.1440370103

Connolly, B., Miller, T., Ni, Y., Cohen, K. B., Savova, G., Dexheimer, J. W. & Pestian, J. (2016). Natural language processing – overview and history. In J. J. Hutton (Ed.), *Pediatric biomedical informatics* (Vol. 10, pp. 203–230). Series Title: Translational Bioinformatics. doi:10.1007/978-981-10-1104-7_11

Felton, J., Mitchell, J. B. & Stinson, M. (2003). Web-based student evaluations of professors: The relations between perceived quality, easiness, and sexiness. *SSRN Electronic Journal*. doi:10.2139/ssrn.426763

Gualtieri, M. & Yuhanna, N. (2016, January 22). Hadoop is data's darling for a reason [Forrester]. Retrieved March 29, 2020, from https://go.forrester.com/blogs/hadoop-is-datas-darling-for-a-reason/

Holmes, R. (2018). University Ranking Watch: Are the rankings biased? Library Catalog: Blogger. Retrieved April 24, 2020, from http://rankingwatch.blogspot.com/2018/02/are-rankings-biased.html

Horton, M. (2019). Simple Random Sample: Advantages and Disadvantages. Library Catalog: www.investopedia.com. Retrieved April 24, 2020, from https://www.investopedia.com/ask/answers/042815/what-are-disadvantages-using-simple-random-sample-approximate-larger-population.asp

Kernel, E. (2019, December 24). Three biggest legal cases about data scraping [JAXenter]. Retrieved April 1, 2020, from https://jaxenter.com/data-scraping-cases-165385.html

Liddy, E. D. (2001). *Natural language processing*. In *Encyclopedia of library and information science* (2nd Edition). NY: Marcel Dekker.

Manning, C. D. & Schütze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, Mass: MIT Press.

Marsh, H. W. (1987, January). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, *11*(3), 253–388. doi:10.1016/0883-0355(87)90001-2

Mengel, F., Sauermann, J. & Zölitz, U. (2019, April 1). Gender bias in teaching evaluations. *Journal of the European Economic Association*, *17*(2), 535–566. doi:10.1093/jeea/jvx057

Mitchell, R. E. (2018). *Web scraping with python: Collecting more data from the modern web* (Second edition). OCLC: on1032828499. Sebastopol, CA: O'Reilly Media.

Mohri, M., Rostamizadeh, A. & Talwalkar, A. (2012). *Foundations of machine learning*. Adaptive computation and machine learning series. Cambridge, MA: MIT Press.

Otto, J., Sanford, D. A. & Ross, D. N. (2008, August). Does ratemyprofessor.com really rate my professor? *Assessment & Evaluation in Higher Education*, *33*(4), 355–368. doi:10.1080/02602930701293405

Otto, J., Sanford, D. & Wagner, W. (2005, June 1). Analysis of online student ratings of university faculty. *Journal of College Teaching & Learning (TLC)*, *2*(6). doi:10.19030/tlc.v2i6.1833

Press, G. (2016, March 23). Cleaning big data: Most time-consuming, least enjoyable data science task, survey says [Forbes]. Retrieved April 28, 2020, from https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#5c8554eb6f63

RateMyProfessors. (2020, March 19). Rate my professors [Ratemyprofessors]. Retrieved March 19, 2020, from https://www.ratemyprofessors.com/

Redman, T. C. (2016). Bad data costs the us $3 trillion per year. *Harvard Business Review*, *22*, 11–18.

Schaal, D. (2019, September 24). Ryanair and expedia settle screen-scraping lawsuits on 2 continents. Library Catalog: finance.yahoo.com. Retrieved April 1, 2020, from https://finance.yahoo.com/news/ryanair-expedia-settle-screen-scraping-173522386.html

Schneider, C. (2016, May 25). The biggest data challenges that you might not even know you have [Ibm]. Retrieved March 29, 2020, from https://www.ibm.com/blogs/watson/2016/05/biggest-data-challenges-might-not-even-know/

Sewwandi, U. (2019). BoW vs TF-IDF in Information Retrieval. Library Catalog: medium.com. Retrieved April 21, 2020, from https://medium.com/@sewwandikaus.13/bow-vs-tf-idf-in-information-retrieval-a325b5e61984

Trstenjak, B., Mikac, S. & Donko, D. (2014). KNN with TF-IDF based Framework for Text Categorization. *Procedia Engineering*, *69*, 1356–1364. doi:10.1016/j.proeng.2014.03.129

Uysal, A. K. & Gunal, S. (2014, January). The impact of preprocessing on text classification. *Information Processing & Management*, *50*(1), 104–112. doi:10.1016/j.ipm.2013.08.006

Wartena, C. & Brussee, R. (2008, September). Topic detection by clustering keywords. In *2008 19th international conference on database and expert systems applications* (pp. 54–58). 2008 19th international conference on database and expert systems applications (DEXA). doi:10.1109/DEXA.2008.120

Wilder-James, E. (2016). Breaking down data silos. *Harvard Business Review*.

Winograd, T. (1972, January). Understanding natural language. *Cognitive Psychology*, *3*(1), 1–191. doi:10.1016/0010-0285(72)90002-3