HUMBOLDT UNIVERSITÄT ZU BERLIN

WIRTSCHAFTSWISSENSCHAFTLICHE FAKULTÄT

INSTITUT FÜR INFORMATIK

MOOC Interrupted
Determination of Disengagement Factors using Sentiment Analysis

MASTERARBEIT

zur Erlangung des akademischen Grades

Master of Science (M. Sc.)

in Wirtschaftsinformatik

eingereicht von:	Laura Gabrysiak Gómez	Gutachter:	Prof. Dr. Stefan Lessman
Matrikel-Nr:	555091		Prof. Dr. Niels Pinkwart
Adresse:	Gerichtstr 12, 13347 Berlin		
Kontaktdaten	gagomezl@hu-berlin.de		
	+49 17680400424		

eingereicht am: 27.09.2016

Abstract

Within the past few years, Massive Online Open Courses (MOOCs), have experienced a sudden rise in popularity, reaching thousands of students across the globe and bridging limitations that higher education has experienced until present. Despite the great potential MOOCs represent to the future of global education, a high drop-out phenomenon (96%) has been regularly reported over the years.

This thesis investigates this phenomenon and uses Sentiment Analysis and other Text Mining techniques to identify topics related to the student drop-out. An en-to-end framework is presented including data extraction and preparation, polarity classification and feature selection. Using the proposed framework, topics related to student dissatisfaction were found, supporting findings from previous studies.

Furthermore, quantitative text metrics were used to identify linguistic patterns among dissatisfied students' reviews, finding significant correlations with course features such as price, rating and polarity.

Contents

Abstract I

List of Figures III

List of Tables IV

Abbreviations V

1 Introduction **1**

1.1 Massive Open Online Courses (MOOCs) 1

1.2 Measuring Opinions: Sentiment Analysis 3

1.3 A Note on Terminology 4

1.4 Thesis Structure 5

2 Related Work and Background **6**

2.1 Research on MOOC Drop-Out Rates 6

2.1.1 Investigating Learner's Behavior 7

2.1.2 Investigating Learner's Opinions 8

2.2 Text Classification and Text Mining 9

2.2.1 Sentiment Analysis (SA) 9

2.2.2 Lexical Analysis Techniques 13

2.3 Publications taken as Reference 15

2.3.1 Adamopoulos (2013) 15

2.3.2 Fang & Zhan (2015) 16

3 Identification of Disengagement Factors **17**

3.1 Thesis Outline and Contribution 17

3.1.1 Thesis Outline 17

3.1.2 Contribution 18

3.2 Underlying Assumptions 18

3.3 Research Questions 19

4 Implementation **21**

4.1 Technical Specifications 21

4.2 Data Extraction 21

4.2.1 Data Source 22

4.2.2 Data Selection Process 22

4.3 Non-Textual Features: Exploratory Data Analysis 23

4.3.1 Course related Variables 24

4.3.2 Student related Variables 25

4.3.3 Insights from the Data 27

4.4 Textual Reviews: Data Preparation 28

4.4.1 Text Normalization 28

4.4.2 POS Tagging and Emotion Sentences Extraction 29

4.4.3 Quantitative Text Analysis (QTA) 31

4.4.4 Data Transformation 32

5 Results & Discussion 33

5.1 Polarity Classification 33

5.2 Disengagement Factors Identification 36

5.3 Summary & Discussion 37

5.4 Research Questions Revision 41

6 Conclusions & Ideas for Future Work 44

Appendices 51

List of Figures

2.1 Literature main research streams 7

2.2 Main techniques and models related to Sentiment Analysis 11

4.1 Selection process of the course dataset 22

4.2 Overview of Coursetalk MOOC providers 24

4.3 Probability distribution of course price data. 25

4.4 Overview of anonymity and completion rate variables 26

4.5 Rating Statistical Characteristics 27

4.6 Probability distribution of text metrics 31

5.1 Comparison of the different polarity classification approaches. . . . 34

5.2 ROC curve for complete dataset 36

5.3 Descriptive terms clustered according to their TF-IDF. 38

5.4 Timeline of Coursetalk according to reviews 39

.1 Block diagram of the implementation 57

List of Tables

1 Correlation table of course related non-textual features 51

2 Correlation table of variables status, polarity and anonymity 51

3 Confusion matrix for the polarity classification 51

4 Summary of most relevant publications 52

5 Overview of all extracted and inferred variables. 53

6 Overview of Coursetalk MOOC providers 54

7 The Penn Tree Bank Project tagset. 55

8 Correlation table of all linguistic metrics 56

9 Relevant factors associated with student dissatisfaction. 56

List of Abbreviations

- MOOC - Massive Online Open Courses
- SA - Sentiment Analysis
- KDD - Knowledge Discovery
- ML - Machine Learning
- NLP - Natural Language Processing
- POS - Part of Speech
- TF-IDF - Term Frequency Inverse Term Frequency
- MI - Mutual Information
- RegEx - Regular Expression
- UX - User Experience
- WC - Word Count
- SC - Sentence Count

1 Introduction

The year 2012 was not only considered to be the potential end of the world by the Maya, but it was also considered to be the "year-zero"and baptized as *the year of the MOOC* by The New York Times [42]. Since then, MOOCs have received a lot of attention by the media and within the educational field, considered by many as a new educational format and disruptive catalyst capable of revolutionizing our current educational system [8]. This thesis investigates the high drop-out phenomenon associated with MOOCs by inferring disengagement factors applying Text Mining (TM) techniques on a self-collected dataset of user-generated MOOC reviews.

1.1 Massive Open Online Courses (MOOCs)

The abbreviation MOOC stands for *Massive Open Online Course(s)* and literally refers to online courses with unlimited access to anyone with a consistent Internet connection that is willing to learn [3]. Within the past few years[1], MOOCs have experienced a sudden rise in popularity and have established themselves as a new educational format [18]. Unlike *conventional* online courses, MOOC show so far particular characteristics as the massive number of MOOC students exceeds by far the size of traditional classes [26]. Open education tools in combination with the latest technological advancements allow professors and educational institutions to reach a massive number of students across the globe. An example reflecting MOOC's impact is the case of MIT's first MOOC *6002x: Circuits and Electronics* which received over 150.000 registrants across 194 countries within the first days [6]. Only 4.6% reached to complete the course and get the certification, however but after all a massive number of 7.000 students obtained a certification from an elite university for no monetary cost else than the invested time. MOOCs therefore allow students to overcome the main limitations that high education has traditionally presented e.g. geographic and economic barriers [45]. Among the most

[1] MOOCs date back to 2008 at their early beginnings however only reached a sudden hype at 2012 [16].

prominent MOOC providers (i.e MOOC platforms) are Coursera, edX, Udemy etc. Table 6 and figure 4.2 show the main MOOC providers as well as the main institutions offering the courses.

The MOOC Drop-Out Phenomenon

However, despite the great adoption MOOCs have experienced, there is still room for improvement [2]. Among the main problems concerning MOOCs, a large drop-out phenomenon has been identified over time [18]. A standard 96% drop-out rate has been confirmed by several investigations across all MOOC disciplines (see [16, 26, 8]). For instance, already in 2012, the average Coursera MOOC enrolled between 40.000 to 60.000 students, from which only 60% returned from the first lecture [26]. This phenomenon has set an alarm within the educational community [18] and considering the great potential of this technology, *it is the duty of the academic community to shed light on the problems of MOOCs, trying to both understand their causes and help open education to achieve its potential and not fail* [2].

However, the discussion around drop-out is not absolute. Creelman (2013) challenges the concept of drop-out metrics and questions if, rather being a sign for deficient (course-) quality, drop-out is just an expression of freedom of choice [11]. Also, Koller et.al (2013) believe retention rates as irrelevant and object that among students engaged *to completing the course*, the retention rate is actually very high [26]. Jordan et.al (2014) argues that quoting enrollment and completion figures of early MOOCs (i.e. 96% drop-out rate) as representative is misleading [23] as the MOOC market is constantly evolving and the demand and offer relation has changed extensively since 2012 [18]. In fact, there isn't yet an existing standardized definition for the exact meaning of *drop-out* so far known. The traditional understanding of this metric is the *percentage of students that enrolled in a MOOC and did not completed it* [9]. However Ho et.al (2013) conclude that certification (completion) as a metric is misleading and even a counterproductive indicator of the MOOC impact [20].

Even if the majority of MOOCs do have completion rates less than 10% of those who enroll [26], the definition of completion rate according to Jordan et.al (2014) should rather be *the percentage of enrolled students who satisfied the courses' criteria in order to earn a certificate* [23]. Simultaneously, Koller et.al (2015) remarks that retention should be always measured within the context of the student's intend and questions if retention is even the right metric to measure the success or quality of a course [25], also because students are already benefiting in many ways by participating in the course even without completing the assessments [26]. Likewise Ho et.al (2013) abet to develop new metrics *far beyond grades and course certification* to better capture MOOC usage [20]. However, MOOC value assessment is a highly subjective and complicated topic [45]. Creelman (2013) suggests that some evaluation criteria could be similar to the traditional syllabus assessment, while some extra metric(s) could be added such as online environment's assessment [11]. MOOC assessment however, goes hand in hand with the student's expectations [18] and these need to be first measured in order to gain a better understanding on this subject. Overall, there is a fundamental need to developing better assessment metrics to understand how learners are interacting with MOOCs [9] and thus to enable MOOC assessment and adapt MOOC design to the student's needs prior to implementation [18].

Despite the large amount of studies forecasting drop-out [5, 3, 54], there is a lack of investigation around the motivation behind this phenomenon even though, only by comprehending the main causes underlying the high drop-out rates, actionable solutions [2] can be provided and the format of the courses can be adapted to maximize its potential [18]. Building on the argument that different MOOC assessment metrics and techniques are needed [9, 20, 6, 18] and given the lack of investigation around the reasons behind the drop-out phenomenon [2], it is the motivation of this thesis to investigate the drop-out phenomenon from the student's perspective. One way of investigating student's perspective and expectation is by analyzing the students' opinions and behavior [44].

1.2 Measuring Opinions: Sentiment Analysis

Sentiment Analysis (SA)[2] is a technique used to classify texts based on its *sentiment*[3] using several methods including Text Mining (TM), Natural Language Processing (NLP), Information Retrieval (IR) or Machine Learning (ML) [13]. In the latter case, SA can be perceived as a text classification task assigning a *sentiment* to an *opinionated* textual utterance[4].

Methodologically, SA represents a great alternative to survey-based studies or any

[2] As already stated at section 1, there is little standardization within the field of Sentiment Analysis. Only few theoretical compilations have been provided by Liu (2012) [30, 32] and Pang (2002) [41].

[3] The concept of *sentiment* can be understood as the attitude, feeling or emotional state of mind [30, 13].

[4] A so called *opinionated* text as the name implies, denominates a text, expression the author's point of view and opinion. Such type of texts are part of our daily lives and range from a product/movie review to a political manifesto.

field investigating human perception [51] as survey based studies yield rather scaled answers on a *pre-selected (biased), limited set of items* [14]. Also questionnaires work by *what you ask is what you get - principle*, which limits a proper statistical analysis in both quality and size [37]. This method was selected as it allows to collect insights at large scale and discover behavioral and perception trends [41]. In fact, SA is vastly applied in diverse domains implying user research such as marketing, socioeconomic studies, but also political sciences and law. For instance SA has been vastly deployed in marketing to measure the perceptions or misconception of products and services [43]. Even the measurement of possible future customers' perception can be measured in order to find opportunities in the market. In politics on the other hand, SA is often used to measure the current perception of current political affairs for example in electoral campaigns such as the US presidential campaign of Barack Obama in 2008 and 2012 which used real time SA on Twitter data and Salesforce to extract key social trends that had to be addressed [53].

SA will be used in this project together with other Text Mining techniques in order to classify the polarity of the investigated user reviews. Section 2.2 will provide a short introduction into this field and later in section (sec.4) will explain how SA is implemented.

1.3 A Note on Terminology

A terminological disagreement in the fields related to this thesis has been observed throughout the compilation of this thesis as perhaps already noticed from sections (sec.1.1) and (sec.1.2). Different *pseudo-standardized* terms referring to the same concept are to be found across the reviewed literature (sec.2). A short elaboration on this matter is therefore needed, for the sake of clarity.

In its summary on SA, Pang and Lee (2008) wrote an extra chapter about the incongruent terminology around his field [41]. The terms: *Opinion* (def. 2.2.1), *Sentiment* and *Subjectivity* are observed to be used very loosely as well as the concepts: *Opinion Mining*, *Sentiment Analysis* and *Subjectivity Analysis* treated basically as synonyms [41]. Within the scope of this thesis, the term *Sentiment Analysis (SA)* will be used as the reference term for the technique and the term *Polarity* as the binomial classification task (sec.1.2).

In the context of MOOCs, on the other hand, a similar phenomenon has been observed. Overall, the concepts *dropout/ drop-out*, *completion/ non-completion*, *attrition*, *retention*, *commitment* and *engagement* are also used very loosely and treated as synonyms. Unlike in the case of SA [41], in this case there is so far known no detailed study about the terminological usage in this field yet. For the scope of this work, the term *drop-out* will be used as reference.

Furthermore, alongside orthographic inconsistencies, the drop-out phenomenon is regarded from both perspectives: the *retention* and the *drop-out* side. For example: *completion/ non-completion* or *engagement/ disengagement*. Both perspectives *might* refer to the same concept however, within the scope of this thesis both perspectives are considered as different problems. This is relevant because e.g. papers within the reviewed literature, taken as a reference for this work (sec.2) have proposed a methodology to identify factors influencing student *engagement* [18, 2] in order to solve drop-out in MOOC. This thesis is focused on identifying *disengagement* factors i.e. factors associated with student drop-out.

1.4 Thesis Structure

The thesis is structured as following: Chapter 2 provides an overview of related work around the drop-out phenomenon in MOOCs and a short theoretical background for some of the techniques used in the implementation. Chapter 3 provides a description of the thesis outline and contribution alongside the main assumptions on which this approach relies. Also the research questions of this thesis are presented. Chapter 4 provides a detailed description of the implementation. Starting with the data extraction and preparation until data transformation and polarity classification. A descriptive data analysis as well as a quantitative text analysis are presented. The results of the polarity classification and the evaluation metrics so as the identification of disengagement factors are presented are presented in chapter 5 along with a discussion around the results. Finally, chapter 6 sums up the findings and proposes ideas for future work. All tables and images can be found in the Appendix section 6.

2 Related Work and Background

This chapter provides an overview on current literature mainly concerning the MOOC drop-out phenomenon and research done around this topic using Sentiment Analysis (SA) and Text Mining techniques. First, a section on MOOC drop-out will be presented, differentiating the reviewed literature according to research done focusing on the student's behavior and projects focused on analyzing the student's opinions. A section on SA will follow, providing a short introduction into this field so as a theoretical background of some of the techniques used for the implementation of this project. Finally, a short description of two studies used taken as reference for this thesis will be provided. Table 4 provides a short summary of some publications relevant to this thesis.

2.1 Research on MOOC Drop-Out Rates

Despite being such an early technology, there is a broad spectrum of research done around MOOCs. In order to simplify the literature review, the different literature patterns observed were visualized (fig. 2.1). Two distinctive research streams can be observed within the literature. One group (top left of figure 2.1) focuses on the research of drop-out from a social perspective [16, 9], using demographic data [20, 19, 17, 8] and analyzing the student's behavior. The other main group (top right of figure 2.1) focuses on the research of drop-out using textual data to analyze the student's opinion.

By identifying the MOOCs target group and their motivation to participate in the courses [18, 20, 19] the first group pursues to understand the adoption of this new technology by society [45]. A critical part of this research stream relies on demographic data which very often is lacking due to the conceptualization of many MOOC platforms [8]. Therefore surveys [8, 16] and alternative methods [44, 17] have to be adopted. Section 2.1.1 will shortly present some insights from this research group. Most of the research done by the second group uses different data sources ranging from clickstream and log data to textual data from forum entries and user reviews. Also, Machine Learning (ML) methods (mostly supervised) are commonly used to predict the student's probability of dropping out the courses [5, 51, 57]. For the scope of this thesis, it will be differentiated between groups based on the nature of their research and the type of data used. On the one side, research using clickstream and log data [5, 3, 54] and on the other hand, research using mainly textual data: either log data, forum entries or textual reviews) [2, 51, 44, 38, 57, 55, 49, 54]. The latter is the main focus for this project. Furthermore, MOOC unrelated review research [40, 41] will be also presented as the it handles the same type of data (user reviews) and methodologies used in this thesis.

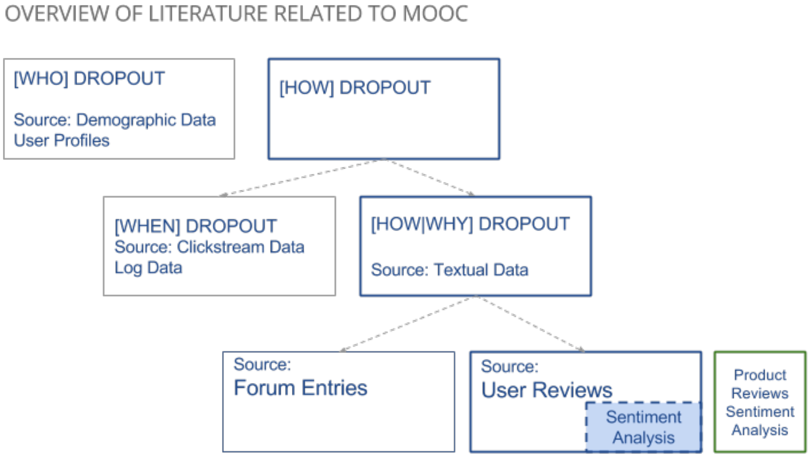

Figure 2.1: Structured overview of the main literature streams. Source: Author's own representation.

2.1.1 Understanding Drop-Out through Learner's Behavior

A great debate is taking place about the social impact of this technology and the *disruptive nature* of MOOCs [45] capable of redesigning our current educational system as we know it. Understanding who is the target group and its main characteristics is the key to adapt MOOCs to the end user's needs [1]. However, this can be very challenging as, in contrast to the massive data recorded about the participation logs, there is a lack of user demographic [17] that can be used to inquire their motivation for participation or dropping-out.

Christensen et.al (2013) ran a survey on 32 MOOCs with a total of 34,779 responses in order to find out exactly these informations. This was one of the first efforts in

[1] Speaking about a *target group* in this context abuts against one of the MOOCs principles (openness). [11] argued that there is *no target group as everybody is welcome as part of the MOOC principles*. However, studies based on demographic data [20, 19] and surveys[8, 9] does show a rather homogeneous group participating in MOOCs [45]. The concept target group refers to this representative group in this context

constructing a dataset to investigate the so called *MOOC population*[2]. The same year and later in 2014, edX published a similar study however in much larger scale [20, 19] resulting in one of the largest surveys of massive open online courses (MOOCs) to date [19] with 597,692 unique users analyzed the first year (2013) and 1.7 Mio observations[3] a year later [19]. Both studies found young professional to be using this technology the most [8, 20]. Given the massive students numbers a large demographic difference was observed[19] although many students were US based (16%-36%). Even though if the enrollment is possible throughout the entire course, the first week was found to be critical as the majority will enroll however around 50% of the students are very likely to drop-out [20].

However these type of studies are rather sparse as useful data is not made available by the MOOC providers. As a result, many (independent) researchers have to enroll in the courses to get access to the data for each course (e.g. [2]), which hinders a large scale quantitative study [51].

Therefore, alternative methods have to be adopted in order to study the student's behavior. Kizilcec et.al (2013), Wong et.al (2015) and Anderson (2014) use forum entries in order to investigate user interactions within the course, finding behavioral patterns called *subpopulations* [24]. Also Anderson (2014) defines different behavior roles (*Bystander, Viewer/ Collector, All-rounder, Solver*) and investigates the chronological development of their posting behavior with regard to their course performance. However this type of research does not provide insight on the reasons why students drop-out. Therefore a goal is to investigate not only the students behavior but also their opinions.

2.1.2 Investigating Drop-Out through Learner's Opinions

One way studying student's opinions is using user-generated textual data and linguistic metrics in order to extract information around the user's behavior and mindset [37, 13]. Furthermore, textual data offers a multidimensional analysis[29], providing not only the textual content but also additional features e.g. analytical, semantic and syntactic features. There are two main streams among the reviewed literature focused on textual analysis. One group uses mostly forum entries (already presented above) and the second user-generated reviews. The latter group, although much smaller, will be of great interest and the baseline for this work.

[2] Although the responses represented approx. (8.5%) and thus suggested a strong selection bias, the study did provide a first insight on the motivation of the students which turned out to be on its larger extent full time employed alumni and not young students as was assumed.

[3] specifically 68 courses, 1.7 million participants, 10 million participant- hours, and 1.1 billion logged events.

When trying to obtain an insight into the users behavior, forum entries and review data have been a commonly used as source to mine opinions [13, 40, 41, 44, 2] as they are created spontaneously and charged with personal assessment (sentiment) [30]. Forum entries on the one hand, have the advantage that are written alongside the course pace [44] so that a chronological documentation is possible, allowing the MOOC designer to quickly spot out and forecast possible dropouts [54]. However, forum entries have the disadvantage of being subject to selection bias as students who post on forums are mostly very engaged with the course (*Solvers*) [4] and are therefore not representative [55] to all students. Also, forum entries are very sensitive to their time frame context as time distorts the data so that e.g. many posts are predominantly negative messages as they are meant as a feedback mechanism not an overall course assessment [54].

Reviews on the other hand, are more extensive and better structured [30]. Also, as they do are purposely written in order to provide an overall assessment [15] they ease the information extraction process[40]. Furthermore, reviews commonly provide additional information useful for the data analysis e.g. the combination of starring (quantitative) and textual information (qualitative)[30]. Nevertheless, aside from the subjectivity and the selection bias phenomenon characterizing natural language [30], textual reviews are also more complex and more difficult to analyze [40]. For instance many complex language phenomena such as negations, sarcasm/irony and humor are very difficult to process [41]. However there are some techniques from Natural Language Processing (NLP) and Text Mining (TM) allowing to extract insights from textual information. Next section will provide an overview of the TM techniques that were used in this project.

2.2 Text Classification and Text Mining

This section will provide a short introduction to some text processing and statistical techniques used in text classification and implemented in this thesis. First a short introduction to Sentiment Analysis (SA) will be provided followed by an overview of some techniques and approaches used in SA and TM in general such as Document Term Matrix and TF-IDF weighting.

2.2.1 Sentiment Analysis (SA)

This section provides an short theoretical introduction[4] to SA starting with a formal framework provided by Liu (2012) [32] followed by some concepts of SA such

[4] For a more detailed theoretical introduction please refer to Liu (2012 and Pang (2002).

as tasks and analysis levels.

Information can be differentiated into *facts* and *opinions* [29]. Contrary to facts, opinions are individual assessments that range in matters of objectivity and polarity [32]. We can define an opinion according the definition provided by Liu (2012) [30, 32, 31][5] An *opinion* is then uttered when a person (*Holder*) expresses his positive or negative view (*Sentiment*) about something (*Target*) at a certain moment (*Time*). These 4 elements are essential and are considered to be the minimal information necessary to successfully recognize and classify an opinion [32]. This definition can be expanded according with the classification task. For instance, this minimal 4-tuple does not differentiate between the place or the gender of the opinion holder because they are not relevant in this case. The reason why we use this format is because it helps to provide a structure to the textual (unstructured) data enhancing the automatic text processing and information extraction.

Furthermore, according to Liu (2012) an *Opinion* can be mathematically defined as the minimal 4-tuple:

$$Opinion = \langle \; \underbrace{T_i}_{\text{Target}}, \quad \mathrm{S}_{ijk}, \quad \mathrm{H}_j, \quad \mathrm{t}_k \; \rangle$$

Where:
T_i: Refers to the *Target* of the opinion
S_{ijk}: Refers to the *Sentiment* about the target
H_j: Refers to the *Holder*
t_k: Refers to the *Time* when the opinion was expressed.

Sentiment Analysis: Tasks and Levels

Sentiment Analysis (SA) refers to a set of classification tasks based on a text's sentiment [41] such as the measurement of a text's *polarity*, (i.e. if the text is negative, neutral or positive) or the identification of a text's specific mood e.g. *happy*, *angry* or*frustrated*. A further task is e.g. the measurement of the text subjectivity, (i.e how subjectively is the text written?) Other authors including [46] also differentiate between the task of extracting texts representing a certain mood (*Opinion Retrieval*) or recognizing if a text contains sarcasm or irony (*Sarcasm and Irony Identification*). For the scope of this thesis however only polarity classification is relevant (and needed). There are three main levels in SA: *document* (or multi-document) level,

[5] Although there are many definitions provided throughout the literature, Liu and Fang have provided a thorough theoretical compilation on the different aspect of the study of Sentiment Analysis. There are several variations of the definition of an Opinion however these all concur mathematically [29]. For further details or comparison refer to [32, 30].

sentence level and *aspect based* SA. The latter surpasses the scope of this thesis thus we will concentrate of the first two levels. There are several Text Mining techniques, for processing *opinions* according to the analysis *levels* and the context of the classification problem [40]. For instance, many investigations using product reviews [15, 58, 48] are rather analyzed at a sentence level or even based on its *aspects* as the topic and the polarity diverges from sentence to sentence [40].

As many classification tasks, SA is defined by the classification method and the feature selection [29]. Following section will provide a short overview of the most important topics around these two elements: Figure (f.2.2) provides a conceptual map with an overview of the most prominent features, feature selection methods and classification models around the SA field[6]

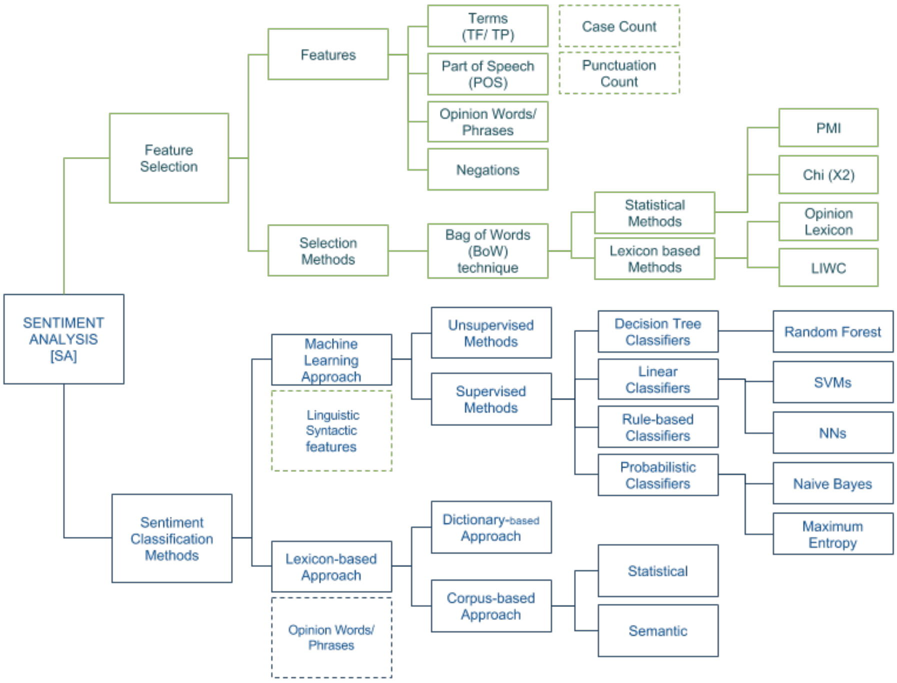

Figure 2.2: Overview of the main techniques and models of Sentiment Analysis. This representation was created by the author inspired by [36] representation.

Every method or feature is selected depending on the context and the nature of the classification task [30] however, neither the methods nor the features are exclusively separated as there are many hybrid methods proposed throughout the SA community [32, 46]. As diagram 2.2 shows, we can differentiate between supervised and unsupervised classification methods. Supervised classification refers to providing the model with a set of manually labeled examples to exemplify how

[6] This conceptual map was inspired by Medhat et.al (2014) and extended with the features and feature selection methods, in an effort to provide a yet more general graphical taxonomic overview of the SA field.

the model (*the decision maker*) should behave [34]. A majority of the reviewed research (sec.1) proposes supervised methods[7] Supervised polarity classification can be thus interpreted is a multi-class learning problem of 3 classes: a text can be either positive, neutral or negative.

$$f(W) = \begin{cases} -1 & \text{if negative polarity, i.e. } Majority(W) \in Dict_{NEG} \\ 0, & \text{otherwise} \\ 1, & \text{if positive polarity, i.e. } Majority(W) \in Dict_{POS} \end{cases}$$

Some approaches [44, 2, 41] leave out the neutral class performing binomial classification, however it has been proven [46, 36] that the introduction of a neutral class enhances the classifier accuracy[8]. Inquiring into the theoretical details of the classification models however surpasses the scope of this thesis as the focus of the implementation is the identification of disengagement factors rather than the polarity classification itself.

Common Features and Feature Selection Methods

Features are key informations that help the model to make the rightful decision therefore the rightful selection is a core aspect in every machine learning task. According to Liu et. al (2012) and Pang et. al (2002), among the most popular features used in SA are Term Frequency (TF), Term Presence (TP), Part of Speech (POS), Negations and Opinion Words [40, 32]. These features will be described alongside the text pre-processing and transformation steps in chapter 4. Moreover, section 5.1 describes the feature selection approach used for the polarity classification. One approach of measuring the TF or TP of Opinion Words is the usage of predefined Opinion Lexicons instead of extracting the features from the text. Next section will provide a short introduction to the implemented Opinion Lexicon.

Opinion Lexicons

Pre-defined sentiment lexicons are a commonly used method (e.g. [31, 52, 2, 13]), and the basis for many SA applications [1]. Several opinion dictionaries and other similar lexical resources[9] available in the Internet include: *Bing Liu's Opinion Lexicon*, the *MPQA Subjectivity Lexicon*, *SentiWordNet*, *LIWC* and *Harvard General In-*

[7] Look up table (tab.4) for an overview

[8] The research proved that the introduction of a neutral class increased the accuracy for Naive Bayes and SVM.

[9] Please refer to [48, 1] for a detailed list of linguistic resources related to sentiment emotion in language

quirer. For the scope of this thesis we make use of *Bing Liu's Opinion Lexicon* [31] and thus these lexical resources won't be covered in this section as it would surpass the scope of this thesis.

Bing Liu's Opinion Lexicon[10] is a compilation of a list of negative and positive words developed over several years starting from 2004 [31, 32]. The corpus was compiled using Amazon reviews [29, 13] and currently includes a total of 6,800 sentiment words (2,006 positive and 4,794 negative words). Although this corpus has several drawbacks such as the indiscrimination of the tokens sentiment strength[11], it does include helpful features such as possible misspellings, morphological variants, slang, and social-media mark-up. Another advantage is that this lexicon is maintained and freely available[31] which allows the replication and rescaling of this project.

2.2.2 Lexical Analysis Techniques

The approach of this implementation is to first detect reviews describing student dissatisfaction e.g based on completion rate or based on review polarity to terms related to student disengagement. The identification of disengagement factors is performed by identifying class-descriptive terms i.e characteristic terms to reviews from students that have dropped out the course or provided negative reviews.

There are several Text Mining techniques for measuring term salience or ranking e.g for Topic modeling such as *Term Frequency - Inverse Document Frequency (TD-IDF), Point-wise Mutual Information (PMI), LDA* . The TF-IDF weighting approach has been selected for this implementation. This section will shortly present the TF-IDF measurement as well as the Document Term Matrix (DTM) technique which is required to calculate the term's salience.

Document Term Matrix (DTM)

This section will introduce the theoretical background of the *TF-IDF* weighting and the *Document Term Matrix* (DTM), implemented later in this project. The abbreviation (*TF-IDF*) stands for *Term Frequency - Inverse Document Frequency* and refers to a quite common technique used in the fields of Information Retrieval and Text

[10] Source:`http://www.cs.uic.edu/~liub/FBS/opinion-lexicon-English.rar`. See files submitted with the thesis [`positive-words.txt`] and [`negation-words.txt`]

[11] Liu's lexicon provide a binomial categorization whereas other resources such as WordNet and SentiWordNet provide a continuous sentiment score. MPQA on the other hand provide binomial categorization with strength categories (weak/strong) Lookup `http://sentiment.christopherpotts.net/lexicon/` for a demo comparing the former mentioned lexical resources based on a token's sentiment.

Mining in order to provide term[12] weighting/ranking in a document collection (i.e. a corpus). This measure can be calculated after converting a corpus into a *Bag of Words* (BoW) model [35] and later into vector space model using the *Document Term Matrix* (DTM). These terms will be shortly presented as they represent the base for all Text Mining techniques implemented in this thesis.

The *Bag of Words* (BoW) model consist in converting a document d_i into a set of all terms contained in it (i.e. $d_i = \{t_1, ..., t_z\}$). The term's order, syntactic function (i.e. POS) or semantic annotation are disregarded in the process [35]. Based on the later, a DTM is finally constructed for the whole corpus (i.e for every document $d_i \in corpusD$). Below a short formal description of a DTM [13] provided by Pang et.al (2002) is presented. Given a Bag of Words model - *BoW* (d_i):

Let ($D = \{d_1, ..., d_i\}$) be a corpus (i.e. a review collection), compound by a predefined set of reviews d_i and ($T = \{t_1, ..., t_m\}$) a predefined set of m terms (i.e. features) that can appear in a review (d_i) e.g.*"course"* or *"great course"*. Each review d_i is then represented by a vector $\vec{d_i}$, where each feature's value is a term weight. The term weight can be either: a binary value (i.e. [0,1] Term Presence - TP), a term frequency value (i.e. Term Frequency - TF) or a TF-IDF (TF)) value. The latter value will be used in the thesis implementation and explained later. Let $n_k(d_i)$ be the number of times that t_k occurs in review d_i, the resulting DTM is a two-dimensional matrix whose rows are the review vectors ($\vec{d_i}$) and columns the term set (T), each element representing a term's TF [40]: $d_i := (n_1(d_i), n_2(d_i), ..., n_m(d_i))$

Term Frequency - Inverse Document Frequency (TF-IDF)

The TF-IDF weighting is a normalized measure of term frequency combining TF and IDF in order to express a terms salience within the class. Moreover, TF-IDF is given by:

$$TF\text{-}IDF(t_m, d_i, D) = TF(t_m, d_i) \times IDF(t_m, D) \tag{2.1}$$

The Inverse Document Frequency (IDF) reflects the relevance of a term t_m in a review within a corpus D [35]. IDF is used to several areas e.g. Information Retrieval to discriminate term (word) relevance and attenuate the effect of too frequent terms [34].

[12] For consistency reasons, the concept *term* is adopted in this thesis as [40] use it in its paper. The term *term* equals the concepts *word* and *token*.

[13] In their paper,[40] Pang et. alrefers to it as the Bag-of-Features Model.

$$IDF(t_m, D) = log \frac{N}{1+ \mid \{d_i \in D : t_m \in d_i\} \mid} \tag{2.2}$$

Where the numerator N refers to the number of document in a corpus D and $\mid \{d_i \in D : t_m \in d_i\} \mid$ refers to the number of documents where the term t_m appears. The denominator is adjusted (i.e $+1$) in order to avoid a zero division in case that $t \notin D$. These techniques will be applied in the project implementation (sec.4.4). First The DTM is used to convert a corpus (i.e a collection of text in this case reviews) into a feature space vector. The TF-IDF weighting is used in order to provide a normalized frequency measurement (sec.2.2.2).

2.3 Publications taken as Reference

Following section will present two investigations that resulted of great interest for this thesis as they are closely related to this thesis in terms of topic and methodology. The first paper (Adamopoulos, 2013) intended to find MOOC engagement factors using user generated reviews. The second paper (Fang & Zhan, 2015) focuses on tackling down the text polarity classification problem using Amazon product reviews. The main approach used in the investigations so as the main drawbacks will be shortly presented.

2.3.1 Adamopoulos (2013)

Adamopoulos pursues to identify determinants affecting *retention* in MOOCs [2]. The study design combined econometric methods, text mining and predictive modeling. In order to identify the retention determinants, Adamopoulos analyzes the opinion of students by measuring the reviews polarity. Only reviews of students who had completed or dropped the course were analyzed with the main focus on the *completed* category [2]. A pre-defined set of features were grouped as a set of characteristics in the following classes: course-, student's, platform's and universities' characteristics. Also a pre-defined set of observed topics were included into the analysis. After measuring the class' respective (average) sentiment score, their respective impact on course completion was measured with a logit linear fitted model. The factors *Professor Assignments* and *Learning Materials* were found to have the largest impact on the completion rate. On the other hand, *Discussion Forum* was identified as having the greatest significant negative impact [2]. The factors were then used in a predictive model using a classifier based on Random Forest [2].

This study was one of the first to combine Text Mining techniques and econometric analysis so far known. It also helped to start a necessary dialogue on the need of investigating the motivation behind the drop-out phenomenon rather than predicting it (sec.1.1). However, the study had some drawbacks that should be addressed. On the one hand, a replication of this study is very difficult due to several reasons. First, the data collection and preparation was partly done manually. The data was reported to be gathered manually via course enrollment and interaction with MOOC students via survey [2]. Also, the final dataset comprised a combination of several data sources (*data triangulation* [2]) that is very difficult to reproduce. Likewise, the data annotation and the identification of the engagement factors was also done manually based on observations[14] done by the author. Secondly, the text pre-processing step was also not described with enough detail so that the data preparation could be replicated. This lead to the motivation of design a scalable and unbiased study easy to be replicated in the future.

Furthermore, Adamopoulos approach targets the completion class and identifies factors influencing completion rate whilst this thesis targets specifically the drop-out class in order to identify factors related to student disengagement. Nevertheless, this paper was chosen as a reference due to the study design and the similarity of the investigated topic and data source used even though if the results cannot be directly compared.

2.3.2 Fang & Zhan (2015)

Fang & Zhan's (2015) paper investigated the text polarity problem [13], performing review and sentence polarity classification on a large dataset of 3 million Amazon product reviews. Even though this paper is MOOC unrelated, it does provide a detailed insight into the methodology used to analyze (product-) reviews. The textual data preparation, including text normalization, are described in great detail, for instance the description of the *ground truth* technique (sec.2.3.2). The models used for the classification were Naive Bayes, SVMs and Random Forest, achieving an average F_1 score of 0.73. The author's also pursued to classifying reviews according to their *real* starring, however this effort showed a very low performance [13]. Overall, this paper was chosen given the detailed description of the polarity classification methodology which was adopted for the implementation of the polarity classification (sec.4).

[14] For the purpose of the study, the author enrolled in several MOOCs and collected the data via surveys (demographic data) and assessment reviews [2].

3 Identification of Disengagement Factors

3.1 Thesis Outline and Contribution

The high importance of measuring student's opinions as a necessary step for a better understanding of the drop-out phenomenon was already explained in chapter (ch.1). It is therefore the motivation of this thesis to investigate the drop-out phenomenon from the perspective of the MOOC students. For this purpose the student's opinion is investigated in order to find patterns related to their behavior. A large dataset of MOOC reviews was self-collected for this thesis from the platform Coursetalk.com[1] for this purpose as dataset was available.

3.1.1 Thesis Outline

In order to investigate disengagement factors from the student's perspective, the identification of disengaged students is essential. This is done by analyzing the students completion rate (drop-out) and the reviews polarity, based on the underlying assumption that students prone to drop-out reflect their dissatisfaction in their reviews [2] (sec.3.2).[2] To identify student's negative reviews a polarity classification is performed using the approach proposed by Fang & Zhan (2015) (sec.2.3.2).

The implementation (sec.4) proceeded as following: First, the review data was extracted with a crawler written for this implementation and then normalized (sec.4.2, sec.4.4.1). Emotion sentences [30] (i.e. text reflecting the student's sentiment) were then extracted and annotated with syntactic information using POS tagging (sec.4.4.2) in order to identify the sentiment carriers. Next, the polarity classification was performed using two lexicon-based methods. Finally, trending terms were identified using class dependent TF-IDF weighting focusing the classes drop-out and negative polarity. The negative reviews were then analyzed with linguistic metrics in order to find linguistic patterns. Additionally, non-linguistic fea-

[1] Source: `www.coursetalk.com`

[2] This assumption is necessary for the project and has been also adopted by earlier studies investigating user motivation with textual data (see [48, 13, 50, 2]). Section (sec.3.2) provides a detailed description of the required underlying assumptions.

tures extracted alongside the textual data were also investigated to find statistical association related to review polarity and completion rate (sec.4.3).

3.1.2 Contribution

Overall, this thesis contributes to MOOC research and (educational) data science with:

- The creation of a crawler in Python for the purpose of collecting user generated MOOC reviews.
- The collection and cleansing of a large MOOC reviews dataset enabling the investigation of MOOC students opinions. The data collection and pre-processing procedures are described in section (sec.4.2).
- The revision of the approaches proposed by Adamopoulos (2013) and Fang & Zhan (2015) (sec.2.3).
- The design and development of an end-to-end automated framework for the identification of factors related to student disengagement, combining techniques from Text Mining (ch.4).
- The identification of factors related with MOOC disengagement using textual review data. The results will be presented in chapter (ch.5).

Furthermore, in order to successfully implement this study framework investigating user behavior in a highly subjective environment such as textual assessments (Opinions, def. 2.2.1), a strict scope limitation so as a set of underlying required assumptions have to be defined. Next section provides an overview of all underlying assumptions adopted for the implementation so as the research questions associated with this thesis.

3.2 Underlying Assumptions

For the successful deployment of this project there are some underlying assumptions that this project is based upon (see section 1). Overall, following assumptions are necessary and core for any investigation related with Opinion Mining and textual data [30]:

Assumption 1: *Opinion Holders express their true opinion (def. 2.2.1) and assessment with their reviews.*

Moreover, building on this idea we also have to assume that:
Assumption 2: *Reviews are genuine and not subject to any manipulation e.g. to influence the rating of a certain course or provider.*

The data source of this work relies on the genuineness of reviews in order to properly measure the users state of mind and assessment. Furthermore, this thesis is focused on investigating factors related to student disengaged based on drop-out and negative reviews as a reflection of student dissatisfaction. Many investigations presented in chapter (ch.3) likewise use this approach. Therefore, for the purpose of this investigation it is assumed that:

Assumption 3: *User disengagement can be measured with drop-out rates.*

Furthermore, if disengagement can be measured with drop-out rates (*Assumption 1*) and it is reflected on the user's review (*Assumption 3*), we can conclude per transitivity that:

Assumption 4: *Drop-out rates are reflected on the user's review and thus the latter can be used for their measurement.*

3.3 Research Questions

After presenting the problem definition (sec.1) and motivation for analyzing textual reviews (sec.2.1) with text mining techniques (sec.2.2.1) as well as having defined the underlying assumptions required, we can proceed with the statement of the research questions that will be taken as reference for the implementation.

Main question: *How can we identify disengagement factors using user generated MOOC reviews in order to gain insights into the MOOC drop-out phenomenon (def. 1.3)?*

The main goal of this thesis is to investigate if and how can MOOC reviews be used to identify specific concepts providing insights into (MOOC) drop-out rates (def. 1.3). It is the assumption that even though Opinions are highly subjective, as they are expressed naturally by the opinion holder (def. 2.2.1), they are expected to be less biased than e.g. a targeted survey study [50] proving to be a good qualitative data source [37]. Taking as reference the investigations of Adamopoulos (2013) and Fang & Zhan (2015) (sec.2.3), it was decided to use sentiment analysis and other text mining techniques (sec.2.2.1) in order to identify relevant terms providing insights into factors associated with MOOC drop-out.

In order to pursue the thesis research question, the TF-IDF technique (sec.2.2.1) will be used to identify salient terms in the class corpus and then measure their association with the targeted class i.e. drop-out using the MI association measure. However, due to the observation of a large class imbalance in the data (sec.4.3.2,5.4), the concept *disengagement* was expanded from focusing only in drop-out rates to focusing into drop-out rates and *student dissatisfaction*. Even though there is yet no concrete definition of a standard course dissatisfaction metric (sec.2), this concept refers in the scope of this thesis to the reviews polarity (i.e if the reviews were positive, neutral or negative). Sentiment Analysis allows us to perform the polarity classification and identify dissatisfied students based on their negative reviews. However, it is yet to proof if:

Sub question 1: *Is there a statistical relation between the drop-out rates and the polarity of MOOC user Opinions?*

This relation has been assumed to be true by several papers [41, 2, 13] processing user reviews however there is so far known no statistical evidence for this. It is investigated therefore if drop-out rates can be linked with student dissatisfaction (based on polarity). The motivation hereby is the ongoing discussion around this topic (sec.1.1).

Also, we know from related studies (sec.2) non-linguistic features have been investigated as part of research in MOOC drop-out. For instance demographic- [8, 20] and click-stream data have been used to investigate social interactions and student behavior patterns [44]. Adamopoulos (2013) (sec.2.3.1) also introduced course related features in addition to textual reviews and analyzed its impact on completion rate [2]. Aside from starring information, non-textual features such as course, student and institution information were extracted with the textual reviews. The following question therefore therefore also arises:

Sub question 2: *What other unknown factors*[3]*, if any, are also influencing drop-out rates?*

The next chapters will outline the implementation of this thesis separating it into two slots: first, the data collection to data transformation steps will be described in next chapter. Chapter (ch.5) outlines the final result so as a discussion and study evaluation.

[3] Unknown refers to not yet investigated factors.

4 Implementation

This chapter provides a detailed description of the implementation of the thesis, starting with the data extraction (sec.4.2) and preparation (sec.4.4) steps. The latter includes the text normalization and POS tagging procedures. Also, two mechanisms for automatic polarity annotation are presented and tested. Finally, the data transformation into the features vector space is described. An exploratory data analysis (sec.4.3) so as a quantitative text analysis (sec.4.4.3) were performed in order to obtain a better understanding of the data. Figure .1 provides a detailed description of the overall implementation process. The final classification results along with the model evaluation and the identification of disengagement factors will be presented in next chapter (ch.5).

4.1 Technical Specifications

This project was implemented with Python and R. The project implementation was visualized with a block diagram (fig. .1) in order to provide a better overview of the project. As shown in the diagram, Python was mainly used for the extraction of the data while R was used for the implementation of the project.

The Python version used was 2.7.10 and the editor PyCharm (Community Edition) JRE: 1.8.0_76-release-b198. while the version used in R was 3.2.2 on a x86-64-pc-linux-gnu (64-bit) platform with the editor RStudio 0.99.491. For the data collection, the Python libraries: [Scrapy] [1] and [BeatifulSoup] were used. For the implementation of the project, the R packages [tm], [openNLP] and [RTextTools] were mainly used. The source code can be found in: https://github.com/gralgomez/moocs_coursetalk_sa/.

4.2 Data Extraction

This section describes the data extraction process of the implementation. First, the data source (Coursetalk.com) will be presented, succeeded by the description of

[1] Lookup at the Scrapy documentation: http://doc.scrapy.org/en/

the undertaken data selection process. The complete implementation is visualized in a block diagram (fig. .1).

4.2.1 Data Source

Coursetalk [`www.coursetalk.com`] is a recommendation platform specialized in MOOCs since 2013 [10]. By February 2016 (when the data was extracted) the platform contained 45,540 total courses. Coursetalk' users were more homogeneous than expected based on previous research done on MOOC users [20, 19, 25]. Most users came from OECD countries: US (76.26%), India (2.67%), UK (1.96%), Canada (1.41%), Israel (1.08%), Germany (0.9%), Ireland (0.77%), Australia (0.74%) and Spain (0.71%) among other countries[2]. Like many other popular recommendation platforms such as *Yelp* or *Amazon*, students can review MOOCs offered by diverse providers such as *edX*, *Coursera*, *Udemy*, etc.

The data used for this project was self collected as there weren't any datasets available that could be used for this type of analysis. None of the major MOOC platforms[3] (table 6) has made the data available so far. In order to acquire data, many (independent) researchers have to enroll in the courses to crawl the data for each course e.g. [2] , which hinders a large scale quantitative study [51]. Coursetalk has been already used in previous studies (see [2]) it was then decided to use this source and ask for cooperation. The data used for this project was crawled on February 2016 with a spider written in Python for the purpose of this thesis.

4.2.2 Data Selection Process

An initial dataset of a total 45,538 courses was crawled from the Coursetalk website in February 2016 and used as a baseline for the extraction of the student's reviews.

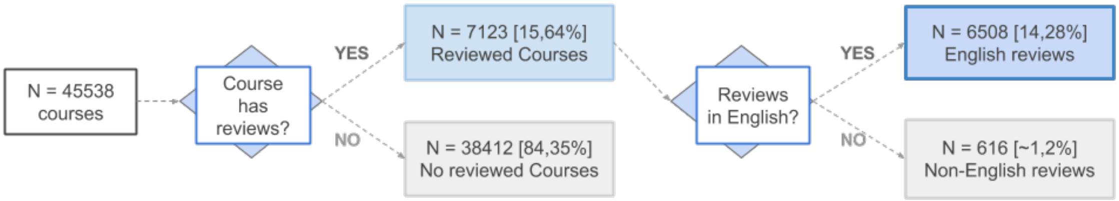

Figure 4.1: Selection process of the course dataset

[2] This figures have been provided by Coursetalk Marketing Department in May 2016. The data was acquired with help of their website analytics. Figures missing: Other countries (12.81%) and Unknown (0.69%)

[3] For example: *edX*, *Coursera*, *Udemy*, etc. Lookup table 6 for a provider overview.

Figure 4.1 represent the selection process undergone by the course dataset. A total 38412 courses (84.35%) has zero reviews reviews and thus were disregarded for the review extraction i.e. 7,123 (15.65%) reviewed courses were kept for further processing. A *language detection* step was performed prior to the review extraction as not all courses or reviews stored at Coursetalk were written in English. This step was necessary as SA is language dependent technique [32]. This step was initially performed with the [`textcat`] R package (`version 1.0-4`) and later via human revision using RegEx[4]. Most common *false friends*[5] to English were Afrikaans, Welsh and German. The package also contains very specific language category profiles such as Scottish that was also mistakenly categorized as English. The [`textcat`] R package uses n-gram categorization, specifically the Cavnar-Trenkle *CT* method[6] [27] was selected by default. After selecting and crawling the target data, a total of 63,806 review observations were extracted however many double entries were observed and as a result 32,656 unique instances were selected as the baseline of this work. It is important to specify that the extracted data was separated in two different modules, handled differently: on the one hand, a set of non-textual (descriptive) features such as the course-, the institution- and the student's characteristics and on the other hand the textual reviews which represented core element of this analysis. Next section provides a description of the non-textual features while section (sec.4.4) will proceed with the data preparation step.

4.3 Non-Textual Features: Exploratory Data Analysis

This section provides a descriptive overview of all non-textual features extracted with the selected review data. The exploratory data analysis intends to gather a better understanding of the target data and its main (statistical) characteristics using mainly descriptive statistics. The target dataset contained a total of 32,656 instances of initially 23 variables. All variables extracted are listed in table 5 and will be described in detail in this section following the table's order (top-down). We start by describing the variables related to the course, followed by the student characteristics. All linguistic variables (table 5) inferred from the textual reviews

[4] Language word frequency lists were used to detect wrongly classified courses and the classification had many false positives

[5] The term *false friends* refers in this context to two or several words from different language alphabets phonetically or orthographically similar (minimal pairs) however different in meaning. In this case, many words of the one language were confused as English due to their orthographic composition.

[6] This method creates an n-gram profile and compares it to a pre-defined (language) category profile matching by the closest distance. The course description was selected as the most suitable variable for the language identification [27].

will be presented in section 4.4.3.

4.3.1 Course related Variables

Alongside the textual reviews, the (course) provider, institution, the characteristics (ranking- and country-), the review number per course as well as the course price information were extracted. Table 6 provides an overview of all MOOC providers along with the quantity of offered courses and the provider-rating which are summarized in figure 4.2. A total of 51 MOOC providers were recorded among them e.g. Udemy being by far the top provider with 79% of the retrieved courses, followed by edX (7.8%), Coursera (5.8%) and Skillshare (3.44%).

The university ranking was also collected following the approach of [2][7]. Almost 21.5% of the total (collected) courses were offered by QS ranked universities, 9% of them were top 100 institutions.[8]. Along with the institution's ranking information, the country of the ranked institution was collected. Most reviewed courses offered by ranked universities were offered by US institutions followed by UK (6.4%), Australia (%) and Canada (3.9%).

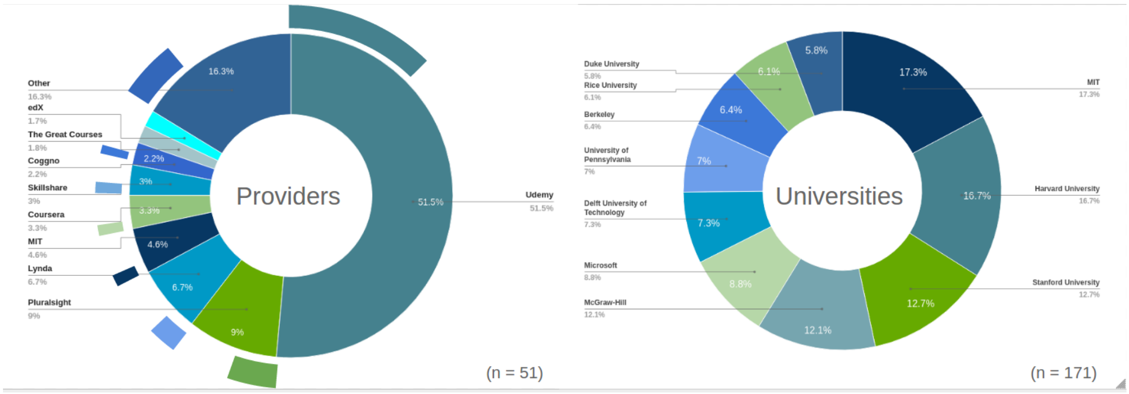

Figure 4.2: Overview of Coursetalk MOOC providers

Price data was also collected and analyzed for the complete course dataset. Almost 25% of the reviewed courses were free of cost and around 57.5% of the rest had an entry cost and 17.4% were based on a monthly fee payment[9]. The price probability distribution of the course with an entry cost is shown in figure 4.3.

[7] As Adamopoulos (2013), the university rank collected from the QS University Ranking (2016). Source: `http://www.topuniversities.com/subject-rankings/2016`

[8] Including in its first ranks world's top universities such as MIT (17.3%), Harvard (16.7%) or Standford (12.7%).

[9] This subset is not visible in the graphic though as the course (time) duration is unavailable, the real cost involved cannot be measured.

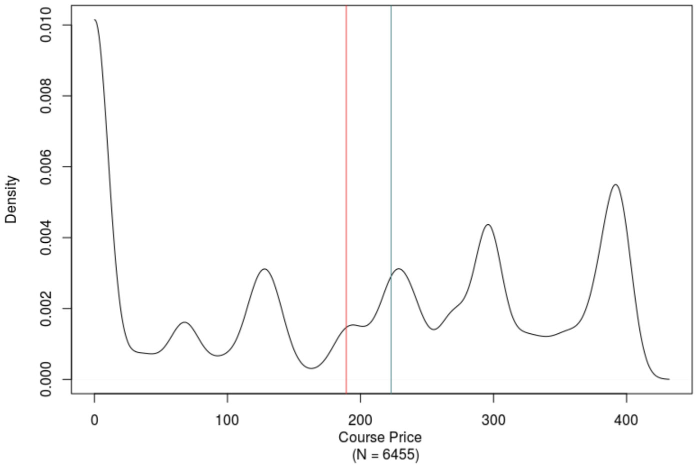

Figure 4.3: Probability distribution of course price data.

4.3.2 Student related Variables

Student related information such as anonymity and completion rate (status) were also extracted. The students' anonymity was inferred based from their ID i.e., if user was logged in the platform or not. Figure 4.4 provides an overview the students' anonymity (left) as well as the students' completion status (right).

Student Anonymity

Unlike expected from MOOCs[10] 4,559 students (27.1%) were anonymous in contrast to the large majority of 12,263 unique students (72.9%) which had an existing profile. Most of users with a profile posted 1 review in average (82.4%) whereas a small group (0.8%) posted ($>$ 10) reviews in average suggesting a *super-poster behavior*[11] [22] as already reported by Huang (2014) and Rose (2014) among other authors (sec.1.1).

Student Status

The student status describes the student's self declared course completion rate when the review was written. This variable is of special interest as it provides direct information insights into the drop-out rates. The different completion rates reported were: *completed*, *taking now* and *dropped*. The group currently enrolled in the

[10] As stated in section (sec.1.1), most of students are passive observers [8] and prone to remain in anonymity [18].

[11] The review/user ratio was also not normally distributed. This was tested with a Shapiro normality test (Shapiro test p ¡2.2e-16).

courses (*taking now*) was not included in the analysis as proposed by Adamopoulos (2013). As figure 4.4 indicates, the dataset shows an alarming class imbalance between the completed (blue) and dropped-out (red) courses. A large majority of students reported to have completed the courses (96.6%) and only (0.5%) to have dropped the course. A class imbalance problem is perceived when in the dataset one class is far larger or more frequent than the other(s) as in this case. This plays a major role aspect for the data classification and analysis performed (sec.5.1).

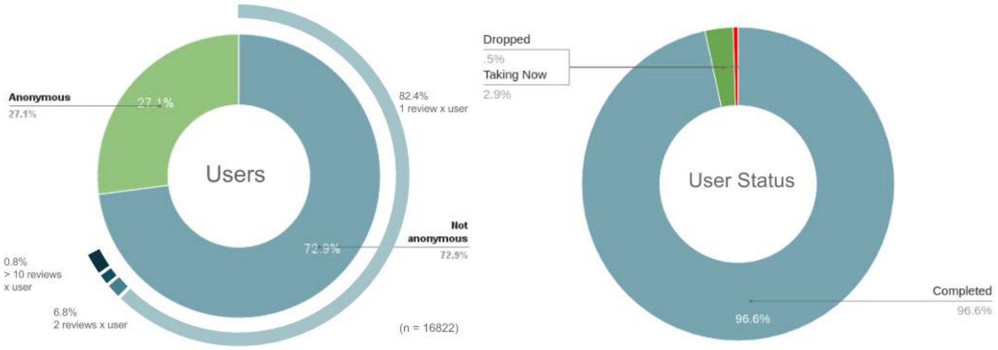

Figure 4.4: Overview of anonymity and completion rate variables

Rating Data

Another variable extracted was the students' rating (i.e. the starring). This variable provides a quantitative assessment measure and can be used as a control variable for the subsequent polarity classification (sec.5.1). The rating information ranged initially from $(0 - 10)$ and was multimodally distributed[12]. The rating was rescaled to a *standard* range of $R_{Std} = (1 - 5)$ [13] adopted by a large number of studies investigating reviews (sec.2). The motivation for rescaling the rating was to ease the comparison of the final results with the rest of the presented studies (sec 2.3). Furthermore, the rescaling helped to structurally increase the class differences[14] and thus increasing the positive- by (46.6%) as well as the negative reviews class by (13.9%).

The rescaling of the rating was performed with a linear transformation. Given two range sets $R_A = (0 - 10)$ and $R_B = (1 - 5)$, so that $f(x) : x \in R_A \to x' \in R_B$:

$$f(x) = \frac{max(R_B) - min(R_B)}{(max(R_A) - min(R_A)) \times (x - min(R_A)) + min(R_B)} \tag{4.1}$$

[12] This can be verified visually (fig. 4.5 and with a Hartigans' dip test for multimodality which proved to be at least bimodal ($non\text{-}unimodal[p - value < 2.2e^{-16}]$)

[13] In this case, this range is considered as standard due to the large number of studies that have adopted it e.g. [13, 21, 31].

[14] Overall, positive reviews increased from 9,506 (29.6%) to 24,438 (76.2%). Negative reviews on the other hand increased from 2,574 (8.0%), to 6504 (20.3%).]

Figure 4.5 shows the different probability distributions of the initial (blue) and the resulting rescaled rating (red). The vertical lines represent the respective mean values. From the graphic it can be observed, that the rating is bimodally distributed. A bimodal (multimodal) rating distribution on reviews has been regularly reported [28] (sec.2) as a reviews characteristic. An explanation to this phenomenon might be the polarized nature of reviews [32], as polarized opinions are mostly the user's main motivation to write a review [40].

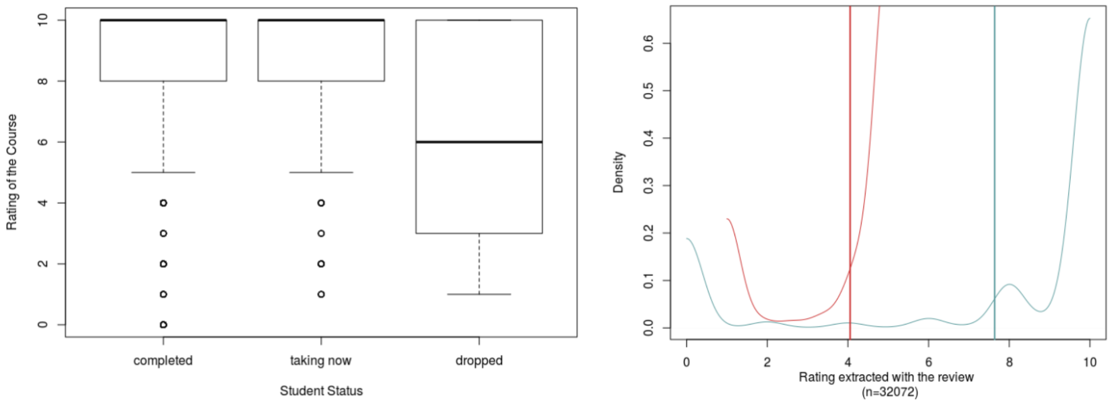

Figure 4.5: Box-plot describing the distribution of status related to rating [Left]. Initial and rescaled rating probability distributions [Right].

4.3.3 Insights from the Data

The course reviews along with their date were extracted[15] Figure 5.4 provides a review timeline, showing a review contribution peak in 2013 and a current drawback. This observation can be interpreted as a general student disengagement supporting the theory of Haggard (2013) (sec.1.1) stating that MOOCs sudden rise is a result of current *technological hype*. This conclusion however, is not statistically justified. The statistical association between the course and student related variables was investigated. Table 1 provides an overview of the respective correlation factors and their statistical significance.

The features *anonymity, status, rating and polarity* were of major interest as they can be used as possible indicators to identify dissatisfied students. As status is a categorical variable, a χ^2 independence test was performed to measure any statistical association. A significant statistical dependency between status and rating so as status and polarity was determined[16] (fig. 4.5) . A significant positive correlation

[15] Even though the course data was not dated which meant a relevant information loss (sec.5), the course reviews were dated.

[16] The χ^2 results were: $\chi^2_{status|rating} = 2773.9, (p < 0.1)$ and $\chi^2_{status|polarity} = 578.54, (p < 0.1)$.

between course price and rating was observed (table 1). A positive correlation between user profiles and completion rates was found (table 2) however also with user profiles (non-anonymous) and negative polarity. Unfortunately, it was not possible to measure a statistical relation between user anonymity and probability of writing a review, as only users who provided a review are collected in the dataset.

4.4 Textual Reviews: Data Preparation

Data preparation is one of the key steps of any Machine Learning task and sets the base for any successful statistical analysis [14]. There is however no strict standard procedure but rather a set of techniques that can be adopted according to the classification task in order to avoid information loss [34]. This section describes in detail the data preparation process performed starting with data and text normalization, POS tagging and finally polarity classification.

4.4.1 Text Normalization

Textual data provided in social media is commonly very noisy [29], containing many spelling, grammatical, and punctuation errors among other language phenomena [32]. In order to improve the accuracy of text classification, clean data is needed [30] and thus a fair amount of data pre-processing is therefore need upfront. Despite the introduced spell-checking step, many orthographic and grammatical anomalies can be only discovered and normalized via RegEx. In the case of SA, as many steps are based on lexical operations (e.g. DTM, TF/ TP and POS), text normalization plays a major role in improving a models accuracy up to $(10\% - 20\%)$ [30].

The reviews were cleaned from any double entries and non-English text. From the originally extracted reviews, 32,523 unique instances were kept. First, all non-alphanumeric characters and diacritical marks (i.e. accents) were removed using RegEx. Also, all apostrophes were resolved e.g. ('ve → have, n't → not, etc.) so as the multiple vowel appearance (e.g. *"veeeery aweeesomeeee"*). A fair amount of slang cleansing was also performed. Abbreviations ([*cuz* | *coz]* → *because*, orthographic alternatives ([*thx* | *thks* | *thanx*] → *thanks*) and abbreviations ([*Idk* | *IDK*] → *I don't know*) were normalized. Furthermore, the lexical differences between US American and British English were normalized to American English. The motivation for this was the large amount (76%) of US based users (sec.4.3).

In order to avoid information loss, punctuation- so as a case scores were inserted. All punctuation as well as caplocks were also counted. Another phenomenon observed throughout the reviews was a large amount of typos, therefore a spell-checking step was also introduced. Spell-checking has been reported to improve the performance of the classifiers up to 20% [13] as it reduces the dimensionality (DTM matrix) which also helps to increase the operational speed. This was performed with the R package [`hunspell`][17].

Negations presence was also scored based on the assumption that a negative polarity is characterized by a large amount of negations [2]. This score however does no measure the scope of the negation [13]. This rather relevant for aspect level SA [30]. After annotating the respective scores, the text was tokenized and converted to lowercase. All punctuation, special characters and numbers were removed after this step. Finally, all stopwords[18] were erased along with sparse terms. Finally, the corpus was stemmed[19] by removing all word inflections and derivations. The removal of stopwords, sparse terms and stemming of the corpus help to reduce the dimensionality of the dataset.

4.4.2 POS Tagging and Emotion Sentences Extraction

After normalizing the textual review, the next step was to identify and extract text relevant to our analysis. As the purpose of this analysis is to differentiate between negative and positive reviews (polarity classification) in order to find a topical pattern, text expressing personal assessment is targeted i.e. we ruled out facts (def. 2.2) and focus on subjective Opinions (def. 2.2.1) which provide the sentiment information. A very common approach to identify Emotion sentences targets sentences containing at least one *emotion word*[20], i.e. a word reflecting sentiment (the sentiment's polarity is not regarded at this step yet).

According to Pang et al.(2002) and Liu et al.(2007), sentiment is mostly reflected by adjectives (JJ), adverbs (RB) and verbs (VB) [12, 41] (denominated *sentiment carriers* from now on)[21] whereas nouns (NN) on the other hand mostly reflect the topic or the Opinion target (def. 2.2.1) [34, 30]. Following the approach used by Fang & Zhan (2015) and Adamopoulos (2013), the Opinion dataset was syntactically

[17] `https://cran.r-project.org/web/packages/hunspell/`

[18] Stopwords refer to very frequent and thus non informative (English) words [32]. In this case, a standardized language dependent stopword lists from the R package [`tm`]

[19] The R package [`SnowballC`] was used hereby.

[20] In the literature we find either the concept opinion word or opinion phrases, sentiment word and subjective word without any specific context. These terms are also used very loosely.

[21] This term has been introduced within the scope of this thesis.

annotated with a Part of Speech (POS) tagger in order to identify the *sentiment carriers*. This step was motivated not only by including syntactical information into the analysis and find the terms providing most (sentiment) information but also to reduce the dimensionality of the feature space vector (sec.2.2.2).

"Awesome course it is fun and a great starting point to web development. If you already know the basics it is a great refreshment. Thanks Rob."(Completed/10)

Above, an example review is provided from the original dataset in order to better understand the followed approach. The (example) review shown is determined by 26 unique words distributed over 3 sentences. According to this approach, the review is first selected as it contains at least one emotion sentence or an emotion word. In fact, all 3 sentences contain emotion words. From the total 9 (positive) emotion words, almost half are sentiment carriers (in this case all adjectives (JJ), t.7)): *"awesome","fun","great","great"*. Therefore, by only selecting the sentiment carriers and comparing their polarity, the overall polarity of the text can be estimated with only 3 unique terms instead of 26.

A POS tagger labels each (previously tokenized) word according to its syntactic function e.g. if the word is an article, noun, adjectives or verb. Table 7 provides an overview of the tagset format used within the scope of this thesis which differentiates 36 different POS. A total of 2,500 unique POS-tagged-terms were annotated. Almost half of the POS tagged terms were sentiment carriers (i.e. JJ/RB/VB, lookup table 7). The identification of the emotion sentences was performed mainly with the R package [`stringr`] and Liu's Opinion Lexicon (sec.2.2.1). The POS tagging step was implemented with a Stanford Maximum Entropy POS Tagger also used[22] by Fang et.al (2015) (sec.2.3). The tagger is provided by the [`openNLP`] and the [`CoreNLP`] R packages and can be implemented with the function `Maxent_POS_Tag_Annotator()`. The tagger uses the Penn Treebank Project tagset format[23], described in table 7.

A total of 1,242 *sentiment carriers* were collected from the 31,474 emotion sentences and grouped into an Opinion dataset[24] which was then then transformed into a feature space vector (DTM sec.2.2.2) and provided to the classifier along with the class polarity tag for training purposes. The data transformation procedure will be described next section.

[22] Adamopoulos (2013) does not provide information on the used POS tagger.

[23] Reference: The Penn Treebank Project `https://www.cis.upenn.edu/~treebank/`

[24] Also referred to in some papers as *Opinion Lexicon*[2, 13] it is not to be confused with Liu's Opinion Lexicon [31].

4.4.3 Quantitative Text Analysis (QTA)

To better comprehend the characteristics of the textual reviews, a quantitative text analysis was performed prior to transforming the Opinion dataset (i.e corpus) into a feature space vector (i.e a DTM). A quantitative text analysis refers to the measurement of descriptive characteristics of a text, mostly at a lexical level [50].

For instance standard metrics are *word count*, *sentence count* but other metrics such as e.g. the number of negative emotion words can be used. Tables 5, and 5 provide an overview of the measured metrics. An average of 3 sentences per review and (39-53) words per sentence were determined. Figure 4.6 provides a description of the probability distribution of word- (gray) and sentence count (black) together the with positive (blue) and negative (red) emotion words ratio. It can be observed from the distributions how negative emotion words are less spread i.e. negative emotion words are less used in this dataset which can be related either with the (rating and polarity) class imbalance observed or with the writing style of the MOOC users.

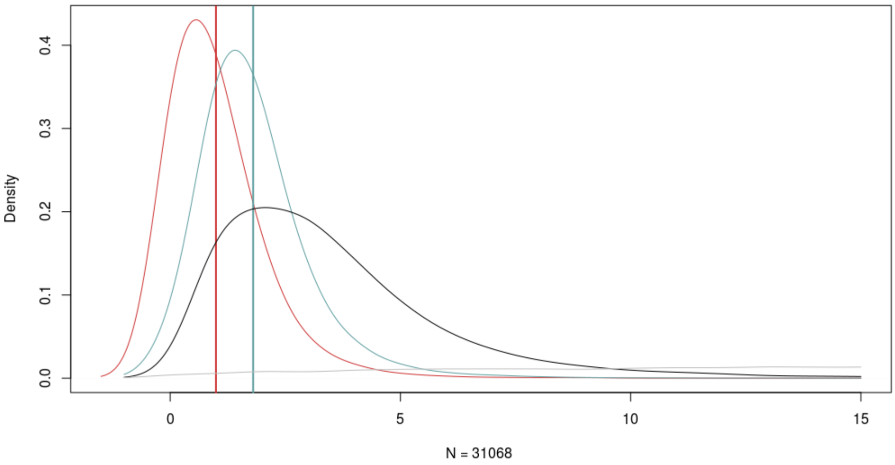

Figure 4.6: Comparison of word and sentence count and positive and negative word ratio

Furthermore, from table 5 it can be observed that there are some outliers including word- (max=1337) and sentence(max=78) count (i.e model features). These outliers are reflected in the word/sentence ratio and emotion words ratio. The emotion words ratio reflect the ratio of positive or negative emotion words per sentence based on Liu's Opinion Lexicon [31]. Table 5 also shows that the mean and the median of negative emotion words are lower than the positive emotion words. Simultaneously negative emotion words ratio has far more outliers.

This is understandable given the subjective nature of polarized opinions. Very dissatisfied users will be prone to write longer, negative reviews in contrast to very satisfied students. Table 8 provides a correlation matrix using Pearson's method across all textual metrics including lexical polarity (equation 5.1). A significant correlation between longer sentences and negative emotion words ratio was determined. The negation ratio on the other hand, refers to the negation count [40] performed in section (sec.4.4). It is also worth noticing that the POS tagged term set is conformed by 1,176 unique nouns (NN) and 1,242 unique sentiment carriers: 643 verbs (VB), 428 adjectives (JJ), and 171 adverbs.

This quantitative text information is relevant for this implementation especially for the data transformation step. When transforming the reviews into the feature vectors. To avoid a DTM with a too high dimensionality (*curse of dimensionality* [33]) outliers with very long sentences should be removed prior the data transformation [40].

4.4.4 Data Transformation

The data was transformed into a feature space vector using a *DTM*. The DTM technique is the ground for all Text Mining techniques implemented in this thesis (sec.2.2.2), providing an overall frequency of every word throughout the reviews and identifies the most salient terms throughout all reviews using the TD-IDF weighting (sec.2.2.

As a result from the quantitative text analysis, reviews with term length beyond the median were sorted in order to every to have a feature vector space containing the approx. same number of dimensions to fit the classifiers [13]. Additionally, all sparse terms were sorted out to reduce the matrix' dimensionality and prevent thus overfitting [35], this phenomenon is better known as the *curse of dimensionality* [33]. As a result a (32,068 x 1,242) vector space matrix was created with a total of 1,242 unique sentiment carriers. All results and the model evaluation, so as insights drawn from the exploratory data analysis as well as the quantitative text analysis will be presented next chapter.

5 Results & Discussion

This chapter provides the final results of the review polarity classification as well as the identification of factors related to student disengagement. First, the classification outcome will be presented alongside it's evaluation. Next, the procedure and final results of the factor identification step will be described in section sec.5.2. The main findings of the implementation will be discussed in section 5.3 and challenges encountered during the implementation will be addressed as well (sec.5.4). Finally, the research questions (sec.3.3) will be revised. The thesis conclusions as well as some ideas for future work are provided next chapter (sec.6).

5.1 Polarity Classification

As a result of the class imbalance data problem already described in section 4.3, the student dissatisfaction concept and therefore the target data has been expanded from collecting reviews of students dropping-out to reviews with a negative polarity. Consequently, a polarity classification has been performed to later carry out with the identification of factors related with student drop-out. Based on Pang et. al (2002) and Fang & Zhan (2015), two different feature selection methods were tested: Opinion Words i.e., sentiment carriers and Term Presence (TP) in Opinion Lexicons. The first approach (Opinion Words) showed a very low performance. Three different models[1] (SVM, Random Forest (RF), Naive Bayes) were tested with a average score of ($F_1 \approx 0.43$). Consequently, the lexicon-based TP was chosen as it showed a much better performance.

According to Pang et. al (2004) (sec.2.3), the review rating is adopted as the *ground truth* [13] as a control measurement and for later evaluation purposes. A manual annotation was not feasible due to the dataset's size. Therefore, a similar approach to Fang & Zhan (2015) was implemented, using a *Bag of Words* (BoW) model and counting the appearance of positive or negative tokens for every sentence (i.e, pos- and neg emotion word ratio) using Liu's Opinion Lexicon [30] for this purpose.

[1] The models were trained with 80% of the data (25,654) and tested with the rest. Subsequently a cross-validation (k=10) was performed.

The overall sentiment score per review has been defined to be equivalent to the average sentiment score of the sentences contained:

$$SS_{Review} = \frac{\sum_{i=1}^{s} SS_{Sent}}{s} \quad , \quad SS_{Sent} = -1 \times \sum_{i=1}^{n} W_{NEG} + \sum_{i=1}^{m} W_{POS} \tag{5.1}$$

Also a Naive Bayes classifier trained on Wiebel's (MPQA) Subjectivity Lexicon [56](sec.2.2.1) was implemented provided by the R package [`sentiment`]. Figure 2.2) shows the different polarity class ratios according to the different classification approaches. The original rating was included in the graphic (light gray) to show the impact of the rating rescaling (sec.4.3). From the graphic it can be observed that the first approach is more sensitive to negative emotion words (7,019 vs. 1,989 predicted negative reviews). A possible explanation for this could be the design of the Opinion Lexicon itself (sec.2.2.1), with twice as much negative- as positive emotion words. Moreover, the inclusion of emotion words with common orthographic variations might have lead to an increase in the recognition of negative emotion words. Even though a spell check step was included in the data preparation (sec.4.4), not every mistake can be ruled out.

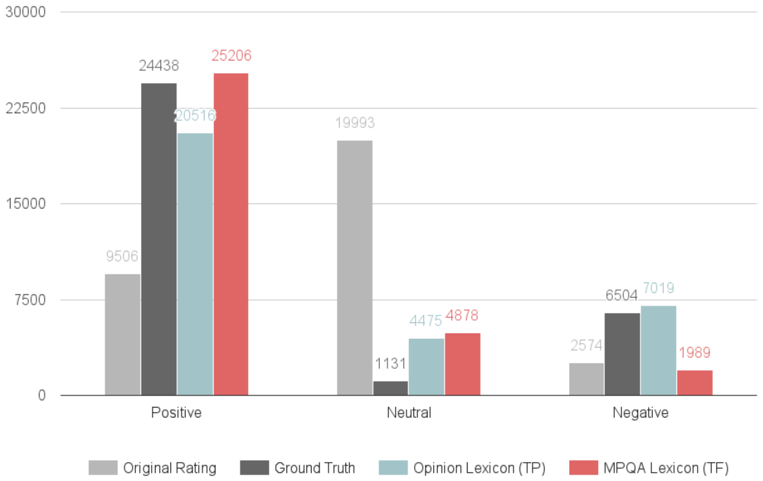

Figure 5.1: Comparison of the different polarity classification approaches.

Classification Evaluation Measures

The evaluation measures used in this work are all based on the F_1 score (equation 5.3), not only because the reference studies (sec.2.3) use this evaluation metric but also due to the observed class imbalance in the dataset (sec.4.3) which the F_1 score manages to overcome[2] using the *Precision* and *Recall* scores[3]:

$$Precision = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FP_i)} \quad , \quad Recall = \frac{\sum_{i=1}^{m} TP_i}{\sum_{i=1}^{m} (TP_i + FN_i)} \tag{5.2}$$

The F$_1$ is defined as following. β refers to the relative importance of *Precision* over *Recall* set to 1, i.e. no difference:

$$F_1 = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 Precision + Recall} \quad , \quad F_1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \ (\beta = 1) \tag{5.3}$$

Table 3 so as figure 5.2 describe the confusion matrix for both classifiers. The first approach showed a score of $F_1 = 0.78$ while the Bayes-based approach $F_1 = 0.86$. This is given due to the its much higher *Precision* (fig. 5.2). The AUC ROC (Area Under the ROC Curve) is another metric used to measure the performance, and to establish dominance relations between classifiers as it is independent of the decision criterion selected and prior probabilities [7] The chosen classifiers' AUC ROC however performed rather poorly (fig. 5.2). Even if the first approach presented a lower F_1 score (blue line), its AUC ROC was larger than the Bayes based approach (red line). On the other hand side, it has been observed that the *True Negative Rate* (TNR)[4] of the first approach ($TNR_1 = 0.78$) exceeded the second approach ($TNR_2 = 0.74$) which confirms the previously observation that the first model based on Term Presence is more sensitive towards negative emotion words. Both models had approx. the same TPR (Recall) $TPR_1 = 0.796$ so as $TPR_2 = 0.795$.

It is possible that the low performance of the ROC curve is related to the distribution difference of the *ground truth*. Even though the rating was rescaled increasing the negative polarity class in size, so did the positive class. A low AUC ROC curve in combination with a high F$_1$ reflects rather than poor performance, a poor model scalability [51] and when used to compare classifiers, it reflects how well the classifiers can differentiate the classes. It was decided to classify the reviews polarity based on the Opinion Lexicon due to the higher ratio shown which is the target class of this project.

[2] Because the F$_1$ score relies on *Precision* and *Recall* (equation 5.2), the performance is measured relative to the different classifications instead of a total class ratio e.g. unlike Accuracy

[3] Precision and Recall measurements for multi-class classification for m=3 classes [-1,0,1] [47].

[4] The true negative rate refers to the models specificity and is given by $TNR = \frac{TN}{TN+FP}$, i.e. the rate of the negative polarity correctly classified

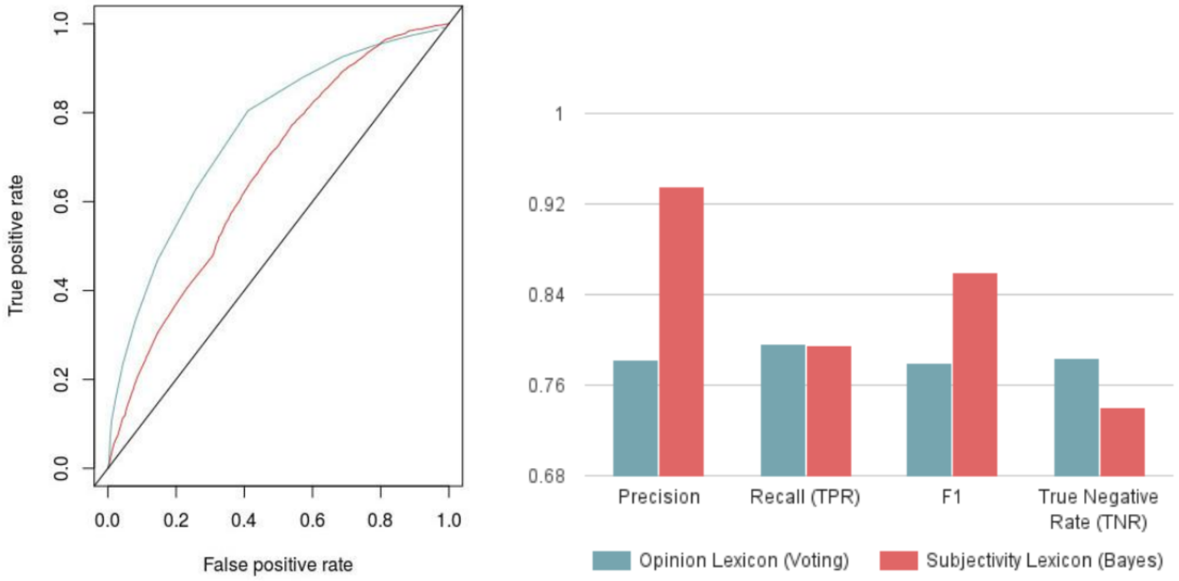

Figure 5.2: ROC curve for complete dataset

5.2 Disengagement Factors Identification

The main purpose of this thesis is to identify factors associated with the drop-out phenomenon (sec.2). As a result of the class imbalance observed in the dataset (sec.4.3), student dissatisfaction was redefined based on the completion rate (i.e. reviews of students who had dropped-out from the courses), and (negative) review polarity. In order to identify negative reviews from the dataset, a polarity classification was performed (sec.5.1) collecting a total of 7,019 negative and 167 drop-out reviews. This section describes the factor identification process followed by the presentation of the results obtained.

Similarly to the previous approach (sec.4.4.2) which selected features based on its syntactic functionality (i.e opinion words/sentiment carriers) (sec.4.4.2), nouns have been chosen for the identification step. A noun does not carry any sentiment [41, 30] but instead, it does provide information around the opinion target [40]. A similar approach is used in Aspect level Sentiment Analysis in order to detect the opinion's aspect [30] collecting a total of . The nouns TF-IDF scores (sec.2.2.2) were ranked in order to obtain the most class-descriptive *terms*.

Results

Table 9 provides an overview of the 30 most class-descriptive terms according to the ranked TF-IDF score. The ranked TF-IDF entries were already pruned in order to sort out common class-descriptive words e.g. *course*.

Many class descriptive terms do match with the observations made by Adamopoulos (2013) e.g. topics related to professor (*instructor, professor, prof*), class materials (*materials, videos, information, content, topic*), and assignments (*assignments*) but also time related terms were observed such as (*weeks, time, hours*). The term *problems* also showed a high salience in both classes. Even though many terms are relevant in both classes and even with the approx. same ranking e.g. *experience*, most of terms ranking is class dependent. Among the drop-out class specific terms the word *drop* and *end* can be found suggesting that students provide explanations on why they dropped in their reviews. Another interesting topic emerging in the drop-out class are the terms *language, english* however these topics have an overall high frequency.

TF-IDF Clustering

In an effort to further investigate the class-descriptive terms, the inter-class TF-IDF scores were clustered[5] (fig. 5.3. Several clusters were identified, some of them inferring semantic relations e.g. (*materials, courses, video*) or (*instructor, questions*). However, given the fact that no semantic information was included in this analysis, it is difficult to establish any causal explanation as how exactly these terms are related to each other. For instance the terms *instructor* and *questions* were associated inferring a possible relation so that this could be interpreted as dissatisfied students expecting the instructor to answer their questions however with no statistical evidence. Therefore the inclusion of semantic information in a future effort is highly recommended.

5.3 Summary & Discussion

One of the main purposes of this thesis is to investigate a methodology capable of identifying factors related to drop-out behavior (sec.2). For this purpose, an approach related to Adamopoulos' (2013) and Fang & Zhan (2015) has been adopted which investigates the MOOC drop-out phenomenon from the students' perspective by analyzing course characteristics and students' reviews. Course starring data (i.e rating) provides limited assessment information [18, 9] and therefore the study of textual reviews is an appealing method for gaining insights into students drop-out behavior (sec.1.1). For this implementation, a dataset was collected and

[5] They were clustered by computing the Euclidean distance. The Euclidean distance is a common measure to cluster matrices. It is defined by the distance between the two vectors $d(x, y) = \sqrt{(x_1 - y_1)^2 + \cdots + (x_n - y_n)^2}$.

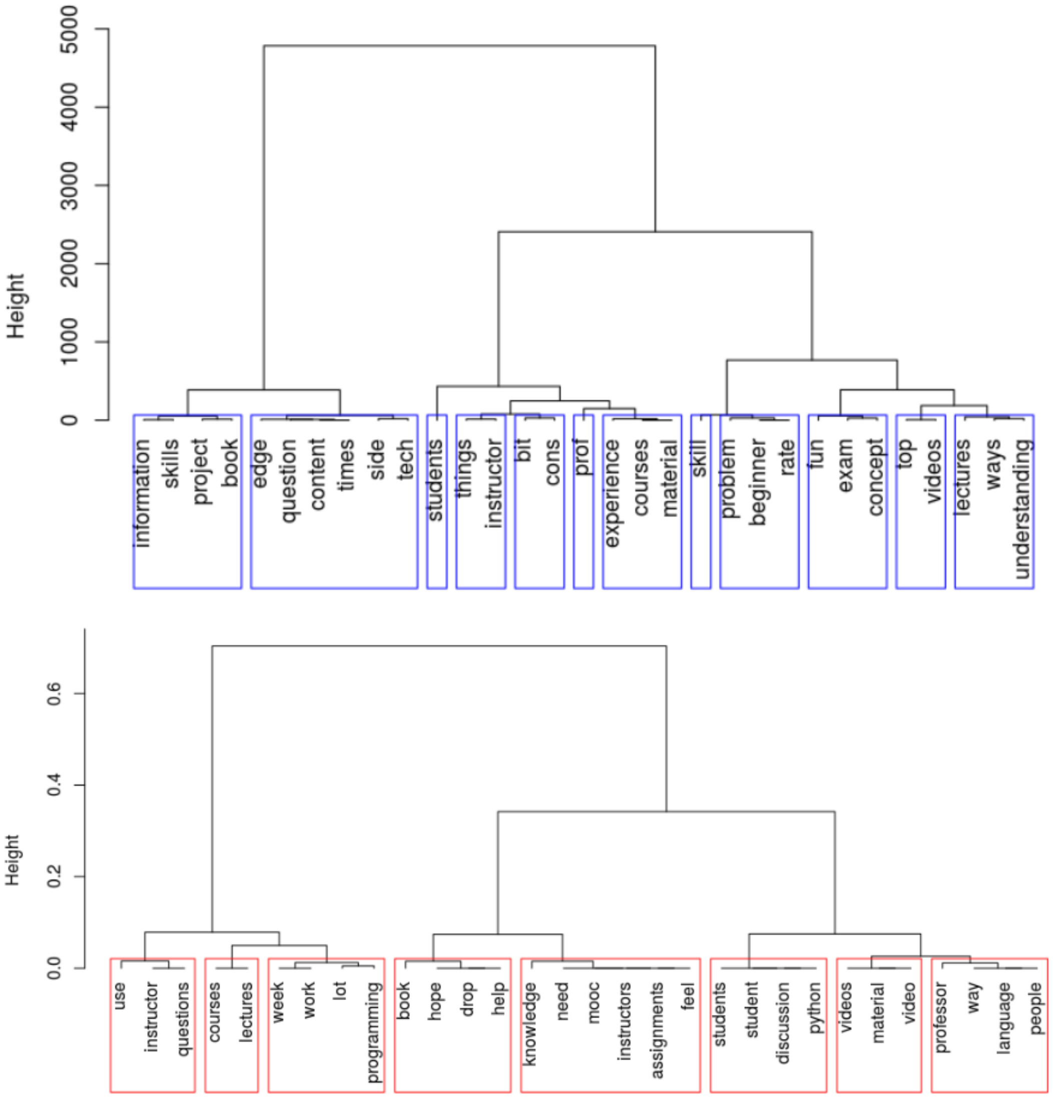

Figure 5.3: Descriptive terms clustered according to their TF-IDF.

normalized (sec.4.2). A scalable end-to-end framework has been provided, enabling a future replication and scaling of this study.

Non-Textual Variables

An insight gained from the descriptive analysis was the data homogeneity (which could explain the observed class imbalance). The majority of students were found to come from the US (92%) which simultaneously infers that the dataset extracted cannot be considered as representative for the global (MOOC-) market. Furthermore, among the MOOC providers, 51.1% of all courses were provided by the platform Udemy (sec.4.3). Many features were also found to be highly skewed or bimodally distributed such as the case of course price (fig. 4.3) and rating.

More importantly, an alarming class imbalance was determined in the student status variable (fig. 4.5), consisting of 96.6% *completed*, 3% *taking now* and 0.5% *dropped*. A large amount of registered users (i.e *not-anonymous*) was observed, however this is understandable in consideration of the dataset's apparent selection bias. Another observation was the reduction of reviews from 2013 on, which supports Christensen et. al (2013) theory explaining the sudden rise of MOOCs as a technological hype (sec.1.1), however this has no statistical evidence.

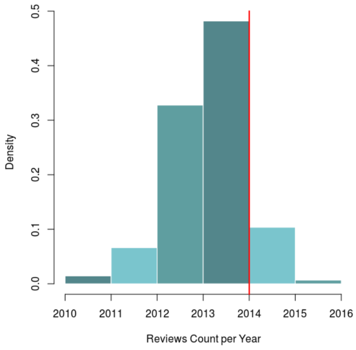

Figure 5.4: Histogram of reviews distributed according to the year (2010 - 2016)

Among the non-textual features, following statistical dependencies were identified:

- A slight negative correlation (-0.036**) between course price and review numbers was determined.
- A positive correlation between provider and price was observed. University ranking and price had also a significant inverse correlation (-0.432**, t.1) which can be interpreted as higher (lower number) ranked universities offering more expensive MOOCs.
- A positive correlation (0.278**) was identified between course rating and course price.
- A significant statistical dependency was determined between registered users and polarity.
- More importantly, a significant statistical dependency could be determined between rating, status ($\chi^2 = 2773.9$) and polarity ($\chi^2 = 578.54$) which serves as foundation for this implementation as it confirms the adopted assumptions (sec.3.2).

Textual Variables

As a result of the observed class imbalance prevailing in the student status, a polarity classification step was introduced in order to identify negative reviews and use them in addition to the drop-out reviews (i.e reviews of students claiming to have dropped-out). Fang & Zhan's 2015) approach was adopted, using lexicon-based methods for automated labeling. Both classifiers showed an average score of $F_1 = 79.5$. Quantitative text metrics were implemented in order to establish language patterns related with negative reviews to best fit the classifiers by sorting out outliers e.g. very long reviews with word count larger than the median (fig.4.6).

Following findings were estimated:

- When comparing the performance of the lexicon-based classification with the ground-truth (i.e the rating) used by Pang et al. (2002) and Fang & Zhan (2015), the question arises if starring is an appropriate control measurement for automatic review labeling. When comparing the class ratios based on the reviews emotions ratio, it can be perceived that (starring) rating tends to be more positive than the overall textual sentiment (fig.5.1) [6]. Fang & Zhan (2015) undertook a related effort to reverse classify reviews to the corresponding starring, however with very low performance [13].
- A significant but moderate negative correlation[7] between longer sentences (i.e sentence length larger than average) and negative emotion word ratio was determined. This advocates for dissatisfied students prone to writing longer reviews than satisfied ones.
- Likewise a significant positive correlation between emotion words and punctuation marks ratio (!?$) so as capslock ratio (i.e sentences written in capslock) was determined. Negative emotion words ratio showed a stronger correlation (punct = 0.549**, case = 0.165**) than the positive ones (punct = 0.512**, case = 0.145**).

The TF-IDF score was employed to identify salient terms in the target classes (drop-out and negative reviews). The TF-IDF (sec.2.2.2 reflects how salient a term is within a class. Similar results to Adamopoulos'(2013) study were achieved even if the study design was not the same (sec.3.2, 2.3.1). Overall it was determined:

[6] A paired t-test was conducted to proof this statement confirming (95%) that the difference between means was not equal to 0. (t = 0.017893, df = 2, p-value = 0.9873)

[7] This was measured with word count (-0.065**) and word per sentence (-0.056**)

- Terms related to *Time, Class Materials, Instructor, Assignments and Difficulty*[2] were observed to be more salient[8] in the drop-out so as the negative reviews classes. These results support the findings of Adamopoulos' (2013) study even though the extracted factors were extracted based on observations whereas however these results were determined based on automated Text Mining techniques which allow a replication.
- Other terms such as *Language, Beginners* and *Fun* were observed to be relevant across all reviews, suggesting to be general relevant topics of interest.
- Terms such as *drop, end* were highly ranked in the drop-out class.

5.4 Research Questions Revision

Sub question 1: *Is there a statistical relation between the drop-out rates and the polarity of MOOC user Opinions?*

The implementation of this effort was made based on several underlying assumptions presented in section 3.2. A significant statistical dependency between rating and completion rate was determined (fig. 4.5) which supports the underlying assumption(s) stated in reviewed studies e.g. [2, 10] and also adopted for this thesis (sec.3.2). Users reflect their opinion and assessment with their reviews and dissatisfied students provide a significant lower rating than satisfied students. However, rating was observed to be skewed and binomially distributed making it questionable if it can be used as the ground truth.

Starring is commonly used as a ground truth as an alternative to manual labeling specially in larger datasets. Even though starring did correlated with the reviews polarity the rating distribution is also very skewed and multimodally distributed. This finding is consistent with previous studies. Therefore the question arises if starring should be considered an objective *ground truth*. Fang & Zhan's attempt of inferring the starring from the review's sentiment showed very low performance.

Sub question 2: *What other unknown factors[9], if any, are also influencing drop-out rates?*

Along with the textual reviews, other features (course-, student and institutions related features were also extracted and analyzed. A positive correlation between dropped status and user anonymity (0.16**) was determined (tab.2). Among the

[8] More salient refers to a higher ranked TF-IDF.

[9] Unknown refers to not yet investigated factors.

frequent topics associated with negative polarity the terms: *Language, English, Time* that can be further investigated in a future.

Main question: *How can we identify disengagement factors using user generated MOOC reviews in order to gain insights into the MOOC drop-out phenomenon (def. 1.3)?*

The chosen approach adopted from Adamopoulos (2013) and Fang & Zhan was successful despite the data limitations observed throughout the study. However a better taxonomy is needed to define the concepts disengaged or dissatisfied students as already exposed in (sec.2). Overall, using the current accessible Text Mining tool available and the algorithms proposed real life data can be processed and used to the benefit of e.g. education by investigating student's opinions regardless the limitations that are bound with text processing.

Model Drawbacks

The presented framework includes several drawbacks and is bounded to limitations that will should be addressed. During the data pre-preparation process many errors were dragged into the analysis e.g. tokenization, stemming, spelling errors or POS tagging errors. Furthermore, despite the great advantage of TF-IDF's easy computation, this score also presents limitations. Based on the BoW model (sec.2.2.2), this approach only analyzes the text at a lexical level i.e., the position of the terms in text is not analyzed, nor its syntactic function i.e., part of speech (POS) or any semantic information. Furthermore, as there is no semantic analysis involved, phenomena such as polysemy or synonymy can not be handled e.g. recognize that the terms *Professor, Instructor* means the same.

A further drawback of this analysis are the textual processing limitations this study is bounded to e.g. the processing of *implicit negative comments*[13]. Implicit negative comments refer to negative reviews including no negative words which are not recognized as they are not included in the respective Opinion Lexicon. Following examples of implicit negative reviews were found in the dataset: *a little light on content., You Sir are a king!.* Neither the words *light* or *king* are included in any Opinion Lexicon[10]. This is one example shows, how cluster based dictionaries or semantic annotation could improve the model. Also, as already mentioned in section 4.4.1, the negation scope could also not be measured.

[10] For this example not only the Opinion Lexicon but all presented lexical resources (see section 2.2.1) were tested.

Data Limitations

There are some data limitations that should be taken into account when analyzing the drawbacks to which this study is bounded. These limitations are worth mentioning in hope they can be taken into consideration for future work:

First of all, there is a strong selection bias present in the extracted data (ch.1). Students writing reviews are not representative of standard MOOC participants, which on the one hand are not only participating mostly passively (*Auditors*) [24, 54] but also have the highest hazard of dropping out the course [26]. In this case, reviews were collected from a *specialized, external* platform which increases the risk of a selection bias. The selection bias of the data was already addressed in Coursetalk' latest report [10] where a self-reported completion rate of 92% was stated. The selection bias was also observed based on anonymity- (5.6%) and the drop-out rate (0.5%) information extracted with dataset (sec.4.3.1). The data however is useful and representative to highly engaged users. These type of users are often used in qualitative research in order to gather insights, known as *extreme users*.

Another limitation shown by the data was that no demographic data available that can be linked to the reviews thus limiting the scope of analysis. Even if the user country rates are available (sec.4.3), they cannot be linked 1:1 with drop-out rates in Coursetalk. Also, the course type is not provided. Coursetalk inserted in June 2016 (after the data collection) a recommendation engine based on the (registered) user which included a course type taxonomy however, the course type information is not provided at the course profile from where the data was crawled. Missing course type data represents a great disadvantage. An effort was undertaken for this thesis to cluster the courses description into topical groups to infer e.g. *course types* with a low performance though. Nevertheless, even when enhancing the data with this information, the type classification is highly subjective and annotator specific as not labeling guideline is provided. Another consideration is if courses should be classified with an multi-class multi-label approach rather than multi-class however this idea can be further developed in future work. There were no guidelines in the website according to which aspects the courses were classified.

6 Conclusions & Ideas for Future Work

This thesis has implemented a scalable automated framework to identify terms associated with drop-out as well as to investigate MOOC reviews at a lexical level. A dataset of 63,806 reviews alongside course-, institution and student features was collected and prepared for the purpose of this thesis (sec.4.4).

A large class imbalance was observed during the implementation and thus Sentiment Analysis was applied to expand the target data by differentiating positive from negative reviews to investigate the latter. For the polarity classification, lexicon-based approaches showed a better performance and thus were selected. Two lexicon-based methods were used, reporting F_1 scores of 0.73 and 0.86 respectively. The F_1 scores are comparable to the performance reported by Fang & Zhan (2015) [13] and a significant difference between starring ratings and textual reviews could be determined.

For the identification of salient terms, the TF-IDF method was applied among reviews of drop-out students and negative polarity. The findings of this study partly concur with Adamopoulos' work even though the methodology or the approach were not the same (sec.3.2,2.3.1). Moreover, by using an automated text processing approach, this implementation identified the same thematic clusters that Adamopoulos collected manually. Terms related to *Time, Class Materials, Instructor, Assignments, Difficulty* were found to be more predominant in the drop-out and negative review classes. Other terms such as *Language* and *Beginners* as well as *Fun* were observed to be relevant across all reviews suggesting them to be topics of interest. The measurement of the actual impact on completion rates surpassed the scope of this thesis and can be pursued in a future study. This automated approach allows for the replication and further development of techniques investigating students' opinions.

Overall we can conclude that user reviews can be used in order to investigate the users' opinions particularly in the context of MOOCs were qualitative data is lacking [8]. The usage of Sentiment Analysis and Text Mining technique offer a toolset that well implemented, represent a great alternative to surveys and quantitative studies, allowing to quantify qualitative studies. The proposed framework can be implemented also in other fields. With the growth of recommendation platforms e.g Amazon, Rotten Tomatoes and forums, it is also an inexpensive way of performing user research.

A future study on how students use MOOCs is necessary in order to gain further qualitative insights in this research area.

Ideas for Future Work

Due to the limited scope of this work, many aspects had to be left out and can be pursued in future work. The course type annotation could not be pursued within this work however it is of great relevance as it provides a more detailed information about drop-out behavior. It would also provide new insights that can be used to improve the MOOCs design. A topic modeling algorithm e.g. Latent Dirichlet Allocation (LDA) can be used in order to cluster yet unknown topical course characteristics.

The inclusion of semantic information could improve the analysis by far e.g. if the system recognizes synonyms such as the *instructor* and *professor*. Also, it was observed that many reviewers explained the reasons to either drop the course or dislike it in a *(if,then)* manner. By investigating this functional phrases, more detailed insights can be derived specifying the students' concrete expectations.

The investigation of external course engagement factors is encouraged for future work. Factors such as Internet connectivity, digital literacy and English language skills should be taken into account since they might have an influence on the MOOC success and student engagement or even represent a completion limitation. Terms related to *Language* were observed to be relevant among the investigated drop-out reviews class.

Bibliography

[1] *Sentiment Symposium Tutorial*, November 2011.

[2] Panagiotis Adamopoulos. What makes a great MOOC? An interdisciplinary analysis of student retention in online courses. In *International Conference on Information Systems 2013*, pages 1–21, 2013.

[3] Quentin Agren. From clickstreams to learner trajectories. Master's thesis, 2014.

[4] Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Engaging with massive online courses. *WWW '14 Proceedings of the 23rd international conference on World wide web*, pages 687–698, 2014.

[5] Girish Balakrishnan and D Coetzee. Predicting Student Retention in Massive Open Online Courses using Hidden Markov Models. *EECS Department, University of California, Berkeley*, 2013.

[6] Lori B Breslow, David E. Pritchard, Jennifer DeBoer, Glenda S Stump, Andrew D Ho, and Daniel T Seaton. Studying learning in the worldwide classroom: Research into edX's first MOOC. *Research & Practice in Assessment*, 8:13–25, 2013.

[7] Nitesh V Chawla. Data Mining for Imbalanced Datasets: An Overview. *Data Mining and Knowledge Discovery Handbook*, pages 853–867, 2005.

[8] Christensen.G, Steinmetz.A, Alcorn.B, Bennett.Amy, Woods.D, and E.J Emanuel. The MOOC Phenomenon: Who Takes Massive Open Online Courses and Why? *Social Science Research Network*, pages 1–14, 2013.

[9] Grainne Conole. MOOCs as disruptive technologies: strategies for enhancing the learner experience and quality of MOOCs. June 2013.

[10] Coursetalk. What Reviews Divulge About Online Education. Technical report, Coursetalk, 2015.

[11] A Creelman. Make hay whilt the sunshines. The corridor of uncertainty. 2013.

[12] Xiaowen Ding and Bing Liu. The utility of linguistic rules in opinion mining. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval SIGIR 07*, page 811, 2007.

[13] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):5, 2015.

[14] Usama Fayyad, G Piatetsky-Shapiro, and Padhraic Smyth. From data mining to knowledge discovery in databases. *AI magazine*, pages 37–54, 1996.

[15] Anindya Ghose and Panagiotis G. Ipeirotis. Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics. *International Journal of Innovative Research and ...*, 23(10):1498–1512, 2012.

[16] Stephen Haggard, William Lawton, Alex Katsomitros, Tim Gore, and Tom Inkelaar. The Maturing of the MOOC. Technical Report 130, Department for Business Innovation & Skills, 2013.

[17] John D. Hansen and Justin Reich. Socioeconomic status and MOOC enrollment. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge - LAK '15*, pages 59–63. ACM Press, 2015.

[18] Sarah Hayes. MOOCs and Quality: A Review of the Recent Literature QAA MOOCs Network. (July), 2015.

[19] Andrew Dean Ho, Isaac Chuang, Justin Reich, Cody Austun Coleman, Jacob Whitehill, Curtis G Northcutt, Joseph Jay Williams, John D Hansen, Glenn Lopez, and Rebecca Petersen. HarvardX and MITx: Two Years of Open Online Courses Fall 2012-Summer 2014. Technical Report 10, 2015.

[20] Andrew Dean Ho, Justin Reich, Sergiy O Nesterko, Daniel Thomas Seaton, Tommy Mullaney, Jim Waldo, and Isaac Chuang. HarvardX and MITx: The first year of open online courses (HarvardX and MITx Working Paper No. 1). *SSRN Electronic Journal*, (1):1–33, 2013.

[21] Nan Hu, Paul A. Pavlou, and Jennifer Zhang. Can Online Reviews Reveal a Product's True Quality?: Empirical Findings and Analytical Modeling of Online Word-of-mouth Communication. In *Proceedings of the 7th ACM Conference on Electronic Commerce*, EC '06, pages 324–330. ACM, 2006.

[22] Jonathan Huang, Anirban Dasgupta, Arpita Ghosh, Jane Manning, and Marc Sanders. Superposter behavior in MOOC forums. 2014.

[23] Katy Jordan. Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, 15(1), 2014.

[24] Rene F. Kizilcec, Chris Piech, and Emily Schneider. Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses. *Lak '13*, page 10, 2013.

[25] Daphne Koller. MOOCS: What Have We Learned? In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 3–3, New York, NY, USA, 2015. ACM.

[26] Daphne Koller, Andrew Ng, Chuong Do, and Zhenghao Chen. Retention and Intention in Massive Open Online Courses: In Depth. June 2013.

[27] Kurt Kornik, Johannes Rach, Christian Buchta, and Ingo Feinerer. *textcat: N-Gram Based Text Categorization Description*, 2016.

[28] Georg Lackermair, Daniel Kailer, and Kenan Kanmaz. Importance of Online Product Reviews from a Consumer ' s Perspective. 1(1):1–5, 2013.

[29] Bing Liu. Sentiment Analysis and Subjectivity. *Handbook of Natural Language Processing*, pages 1–38, 2010.

[30] Bing Liu. *Sentiment Analysis and Opinion Mining*, volume 5. 2012.

[31] Bing Liu and Minqing Hu. Mining Opinion Features in Customer Reviews. *19th national conference on Artifical intelligence*, pages 755–760, 2004.

[32] Bing Liu and Lei Zhang. A survey of opinion mining and sentiment analysis. *Mining Text Data*, 9781461432234:415–463, 2012.

[33] Kamber M., Han J., and Pei J. *Data Mining: Concepts and Techniques*. The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, 2 edition, 2006.

[34] Christopher D Manning and Hinrich Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, 1999.

[35] Christopher.D Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[36] Walaa Medhat, Ahmed Hassan, and Hoda Korashy. Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, 5(4):1093–1113, 2014.

[37] Matthias R. Mehl. Quantitative Text Analysis . *ResearchGate*, (January 2006), 2006.

[38] D.F.O. Onah, J.E. Sinclair, and R. Boyatt. Exploring the Use of MOOC Discussion Forums. November 2014.

[39] Bo Pang and Lillian Lee. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[40] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002.

[41] Po Pang and Lillian Lee. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[42] Laura Pappano and New York Times. The Year of the MOOC, 2012.

[43] Sharan Kumar Ravindran and Vikram Garg. *Mastering Social Media Mining with R*. PACKT Publishing, 2015.

[44] Carolyn Penstein et. al Rose. Social factors that contribute to attrition in MOOCs. In *Proceedings of the first ACM conference on Learning @ scale conference (L@S '14)*, pages 194–198. ACM Press, 2014.

[45] E. Russell, D.M., Klemmer, S., Fox, A., Latulipe, C., Duneier, M., and Losh. Will Massive Online Open Courses (MOOCs) Change Education? *CHI '13 Extended Abstracts on Human Factors in Computing Systems. Paris, France: ACM,*, pages pp. 2395–2398, 2013.

[46] Jesus Serrano-Guerrero, José A. Olivas, Francisco P. Romero, and Enrique Herrera-Viedma. Sentiment analysis: A review and comparative analysis of web services. *Information Sciences*, 311:18–38, August 2015.

[47] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437, 2009.

[48] Carlo Strapparava. Emotional Language in Persuasive Communication. Master's thesis, 2015.

[49] Glenda S Stump, Jennifer Deboer, Jonathan Whittinghill, and Lori Breslow. Development of a Framework to Classify MOOC Discussion Forum Posts: Methodology and Challenges. 2013.

[50] Yla R Tausczik and James W Pennebaker. The Psychological Meaning of Words : LIWC and Computerized Text Analysis Methods. 2010.

[51] Colin Taylor. *Stopout Prediction in Massive Open Online Courses*. PhD thesis, Massachusetts Institute of Technology (MIT), 2014.

[52] Peter D Turney. Thumbs up or thumbs down? Semantic Orientation applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, (July):417–424, 2002.

[53] Jing Wang, Clement T. Yu, Philip S. Yu, Bing Liu, and Weiyi Meng. Diversionary comments under political blog posts. *Proceedings of the 21st ACM international conference on Information and knowledge management - CIKM '12*, V(June):1789, 2012.

[54] Miaomiao Wen and Carolyn Penstein Rose. Identifying Latent Study Habits by Mining Learner Behavior Patterns in Massive Open Online Courses. *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, pages 1983–1986, 2014.

[55] Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rose. Linguistic Reflections of Student Engagement in Massive Open Online Courses. 2014.

[56] Janyce Wiebe, Cem Akkaya, Alexander Conrad, and Rada Mihalcea. Improving the Impact of Subjectivity Word Sense Disambiguation on Contextual Opinion Analysis. Conference on Computational Natural Language Learning (CoNNL 2011)., 2011.

[57] Jian-syuan Wong, Bart Pursel, Anna Divinsky, and Bernard J Jansen. An Analysis of MOOC Discussion Forum Interactions from the Most Active Users. 2015.

[58] Qiang Ye, Rob Law, and Bin Gu. The impact of online user reviews on hotel room sales. *International Journal of Hospitality Management*, 28(1):180–182, 2009.

Appendix

	course_rating	university_rank	review_num	course_price
course_rating	1	-0.039**	0.006	0.278**
university_rank	-0.039**	1	0.002	-0.432**
review_num	0.006	0.002	1	-0.036**
course_price	0.278**	-0.432**	-0.036**	1

Table 1: Correlation table of course related non-textual features.$p < 0.01$**, $p < 0.05$*

	profile	stat.drp	stat.tkn	stat.cpm	pol.pos	pol.neutr	pol.neg
profile	1.00	-0.040**	-0.091**	0.099**	-0.212**	-0.020**	0.218**
stat.drp	-0.040**	1.00	-0.012*	-0.385**	-0.031**	0.101**	0.01
stat.tkn	-0.091**	-0.012*	1.00	-0.918**	0.055**	0.058**	-0.069**
stat.cpm	0.099**	-0.385**	-0.918**	1.00	-0.039**	-0.093**	0.060**
pol.pos	-0.197**	-0.051**	0.040**	-0.016**	0.977**	-0.342**	-0.902**
pol.neutr	-0.020**	0.101**	0.058**	-0.093**	-0.132**	1.00	-0.096**
pol.neg	0.218**	0.01	-0.069**	0.060**	-0.974**	-0.096**	1.00

Table 2: Correlation table of variables status, polarity and anonymity .$p < 0.01$**, $p < 0.05$*

	C_1 Op.Lex	C_2 Subj.Lex
True Positives (TP)	14007	16745
False Negatives (FN)	3613	4278
False Positives (FP)	3877	1139
True Negatives (TN)	1063	398
Precision	0.783	0.936
Recall (TPR)	0.796	0.795
F_1	0.78	0.86
True Negatives Rate (TNR)	0.784	0.741

Table 3: Confusion matrix for the polarity classification

Author	Year	Title	Description	(SS) Method	Classifier
Adamopoulos	2015	What makes a great MOOC? An interdisciplinary analysis of student retention in online courses	Identification of MOOC retention determinants. Study combines econometric analysis, text mining and predictive modeling. Analysis of the reviews and scoring them with sentiment analysis mechanism. Salient concepts were later used as independent variables in a regression.	Similar to Hu and Liu 2004	Random Forest (for multiclass) [predictive model]
Coursetalk Report	2015	What Reviews Divulge About Online Education	Report portraying the platform's MOOC student's reviews and their polarity. The report investigated the correlation between the course's rating and the most common ″latency key-words″ which were turned into categories.	Not reported	Not reported
Fang, Zang	2015	Sentiment analysis using product review data	Identification of Amazon review's polarity. The authors use Amazon reviews and focus into the sentiment polarity categorization problem. They propose a framework that deals with negation and run categorization on sentence and review level. In addition, a thorough presentation of the sentiment analysis problem is also depicted.	Normalized sentiment score (SS) with starring	[Sentence level: 1-manual, 2-machine labeled], [Review level: stars] — SVM, Naive Bayes (F1) >Random Forests (F1)
Rose et al	2014	Sentiment Analysis in MOOC Discussion Forums: What does it tell us?	Mining collective sentiment from forum posts in a Massive Open Online Course (MOOC) in order to monitor students' trending opinions towards the course and major course tools, such as lecture and peer-assessment. A correlation between the sentiment ratio measured based on daily forum posts and number of students who drop out each day is observed.	Survival modeling	-
Ghose et.al	2012	Estimating the Helpfulness and Economic Impact of Product Reviews: Mining Text and Reviewer Characteristics	Identification of product review determinants on helpfulness and impact. Study combines econometric analysis, text mining and predictive modeling. Analysis of the reviews are based on text features identifying text features with high predictive power which are later scored with a regression.	Normalized sentiment score (SS) with starring	Random Forests
Bing Liu	2012	Sentiment Analysis and Opinion Mining	A thorough theoretical overview on the sentiment analysis problem with insight into the (problem) definition and difficulties eg. negation.	-	-
Liu and Zhang	2012	A survey of opinion mining and sentiment analysis	A summarized introduction of Sentiment Analysis with focus on aspect based Sentiment Analysis. Also, on overview of methodologies used by the literature. Special aspects such as sentiment shifting and sarcastic sentences are presented.	-	-
Hu and Liu	2004	Mining and summarizing customer reviews	Applied Sentiment Analysis to summarize product reviews. A sentiment score technique is presented with very effective results.	Summed up the sentiment scores of all sentiment words in a sentence or sentence segment	-

Table 4: Summary of relevant related publications

	Variable Name	min	max	mean	median	sd	skew
Course related variables	Title	-	-	-	-	-	-
	Course Provider	-	-	-	-	-	-
	Course Institution	-	-	-	-	-	-
	Institution Rank	1	800	38	83.8	2.98	7.63
	Institution Country	-	-	-	-	-	-
	Review Number	1	2780	9.82	4	45.69	41.58
	Course Price	0	402	189.31	223	187.34	-0.05
Student related variables	Student Name	-	-	-	-	-	-
	Student Anonymity	0	1	0.17	0	0.38	1.73
	Student Status	-1	1	0.96	1	0.21	-6.12
	Status Dropped	0	1	0.01	0	0.07	13.75
	Status Completed	0	1	0.97	1	0.18	-5.13
	Status Taking Now	0	1	0.03	0	0.17	5.62
	Course Rating	0	10	8.33	9	2.03	-2.79
	Course Rating (Norm)	1	5	4.05	5	1.56	-1.31
Date	Date (Year)	2010	2016	2013.61	2014	0.85	-0.5
	Date (Month)	1	12	5.87	5	3.3	0.33
Review related variables	Word Count	0	1334	52.8	39	53.61	4.13
	Sentence Count	1	78	3.58	3	2.88	3.88
	Negations Count	0	31	0.98	0	1.56	3.42
	Punctuation Count	1	501	5.23	4	7.49	21.64
	Caplocks Count	0	60	1.17	1	1.16	20.08
	Polarity (Ground Truth)	-1	1	0.56	1	0.81	-1.35
	Polarity (Opinion Lexicon)	-1	1	0.42	1	0.83	-0.92
	Polarity (MPAQ)	-1	1	0.72	1	0.57	-1.95

Table 5: Overview of all extracted and inferred variables.

Provider	Course Num	Review Num	Review/ Course	Average Price
Udemy	5005	41539	8.30	237.59
edX	510	6229	12.21	6.38
Coursera	375	5670	15.12	0.00
Skillshare	224	5209	23.25	0.00
Udacity	44	225	5.11	0.00
StraighterLine	43	873	20.30	280.51
Code School	38	2686	70.68	239.00
Treehouse	37	503	13.59	198.00
FutureLearn	33	43	1.30	0.00
The Great Courses	25	25	1.00	185.80
EdCast	24	35	1.46	32.92
Open2Study	18	34	1.89	0.00
Stanford Online	15	240	16.00	0.00
OpenLearning	14	42	3.00	26.80
Edraak	7	245	35.00	0.00
Textile Learning	7	14	2.00	66.00
Iversity	6	8	1.33	0.00
Smartly	6	17	2.83	0.00
Alison	5	5	1.00	0.00
Codecademy	5	49	9.80	0.00
Lynda	5	5	1.00	138.00
NovoEd	5	7	1.40	0.00
Oxford Royale Academy Prep	5	12	2.40	354.40
Sophia	4	4	1.00	249.00
tuts+	4	5	1.25	72.00
First Business MOOC	3	8	2.67	0.00
MRUniversity	3	4	1.33	0.00
openHPI	3	3	1.00	0.00
Pluralsight	3	3	1.00	239.00
Sally Ride Science	3	7	2.33	374.00
Thinkful	3	6	2.00	285.67
Canvas Network	2	2	1.00	0.00
CareerFoundry	2	4	2.00	0.00
International Writing Program	2	3	1.50	0.00
SchoolKeep	2	2	1.00	312.00
Coder Manual	1	5	5.00	302.00
Coggno	1	1	1.00	191.00
Coursetalk	1	5	5.00	0.00
Craftsy	1	2	2.00	0.00
DataCamp	1	2	2.00	198.00
Ed2Go	1	1	1.00	339.00
Filtered	1	8	8.00	45.00
FX Academy	1	1	1.00	0.00
GoSkills	1	1	1.00	279.00
IAI Academy	1	1	1.00	0.00
k-12	1	1	1.00	0.00
Khan Academy	1	1	1.00	0.00
MIT	1	1	1.00	0.00
MongoDB University	1	1	1.00	0.00
One Million by One Million	1	4	4.00	5.00
Openlearning	1	42	42.00	26.80
MED	3	6	2.00	0.00
AVG	127.57	1251.82	6.77	91.82
TOTAL	6506	63843		

Table 6: Overview of Coursetalk MOOC providers along with provider rating and course number.

i	Tag	Description
1.	CC	Coordinating conjunction
2.	CD	Cardinal number
3.	DT	Determiner
4.	EX	Existential there
5.	FW	Foreign word
6.	IN	Preposition or subordinating conjunction
7.	JJ	Adjective
8.	JJR	Adjective, comparative
9.	JJS	Adjective, superlative
10.	LS	List item marker
11.	MD	Modal
12.	NN	Noun, singular or mass
13.	NNS	Noun, plural
14.	NNP	Proper noun, singular
15.	NNPS	Proper noun, plural
16.	PDT	Predeterminer
17.	POS	Possessive ending
18.	PRP	Personal pronoun
19.	PRP$	Possessive pronoun
20.	RB	Adverb
21.	RBR	Adverb, comparative
22.	RBS	Adverb, superlative
23.	RP	Particle
24.	SYM	Symbol
25.	TO	to
26.	UH	Interjection
27.	VB	Verb, base form
28.	VBD	Verb, past tense
29.	VBG	Verb, gerund or present participle
30.	VBN	Verb, past participle
31.	VBP	Verb, non-3rd person singular present
32.	VBZ	Verb, 3rd person singular present
33.	WDT	Wh-determiner
34.	WP	Wh-pronoun
35.	WP$	Possessive wh-pronoun
36.	WRB	Wh-adverb

Table 7: The Penn Tree Bank Project tagset.

	wc	sentc	wps	pct.excl	punct	case	neg	neg.ratio	pos.wds	neg.wds	tot.op.wds	lex.pol.scr	pol
wc	1	870**	0.373**	0.072**	0.625**	0.201**	0.723**	0.213**	0.805**	0.843**	0.902**	-0.065**	-0.165**
sentc	0.870**	1	0.021**	0.133**	0.642**	0.178**	0.635**	0.059**	0.749**	0.731**	0.810**	0.005	-0.104**
wps	0.373**	0.021**	1					0.481**	0.247**	0.288**	0.293**	-0.056**	-0.140**
pct.excl	0.072**	0.133**	-0.061**	1	0.275**	0.045**	0.050**	-0.020**	0.094**	0.046**	0.077**	0.058**	0.037**
punct	0.625**	0.642**	0.095**	0.275**	1	0.152**	0.462**	0.079**	0.512**	0.549**	0.580**	-0.057**	-0.097**
case	0.201**	0.178**	0.100**	0.045**	0.152**	1	0.156**	0.066**	0.145**	0.165**	0.170**	-0.027**	-0.038**
neg	0.723**	0.635**	0.269**	0.050**	0.462**	0.156**	1	0.620**	0.532**	0.600**	0.620**	-0.096**	-0.159**
neg.ratio	0.213**	0.059**	0.481**					1	0.109**	0.163**	0.149**	-0.070**	-0.116**
pos.wds	0.805**	0.749**	0.247**	0.094**	0.512**	0.145**	0.532**	0.109**	1	0.670**	0.912**	0.388**	0.151**
neg.wds	0.843**	0.731**	0.288**	0.046**	0.549**	0.165**	0.600**	0.163**	0.670**	1	0.915**	-0.425**	-0.416**
tot.op.wds	0.902**	0.810**	0.293**	0.077**	0.580**	0.170**	0.620**	0.149**	0.912**	0.915**	1	-0.025**	-0.148**
lex.pol.scr	-0.065**	0.005	-0.056**	0.058**	-0.057**	-0.027**	-0.096**	-0.070**	0.388**	-0.425**	-0.025**	1	0.700**
pol	-0.165**	-0.104**	-0.140**	0.037**	-0.097**	-0.038**	-0.159**	-0.116**	0.151**	-0.416**	-0.148**	0.700**	1

Table 8: Correlation between all linguistic metrics (QTA)

	Drop-Out		Negative Pol		All	
	TF-IDF	Desc.Word	TF-IDF	Desc.Word	TF-IDF	Desc.Word
1	143.866299	drop	5990.6756	video	13.6122143	place
2	134.590688	app	5111.0743	son	3.4203332	novice
3	111.888242	age	5080.5044	cway	2.4005488	comprehend
4	98.634724	interest	4639.648	info	2.1662533	wow
5	94.284283	end	4579.2052	program	1.9513649	differences
6	89.969522	art	4564.1087	student	1.638018	resources
7	89.969522	man	4413.4627	sign	1.5828838	tom
8	89.969522	week	4405.9458	lecture	1.5108765	instance
9	81.452413	lecture	4151.2649	look	1.4637971	event
10	81.452413	thing	4069.2492	work	1.3978741	optimization
11	77.25302	form	3824.3664	dev	1.3416086	emails
12	77.25302	video	3816.9737	clot	1.3310227	errors
13	77.25302	work	3802.1932	part	1.2239347	interact
14	73.095238	instructor	3787.4194	want	1.069888	enroll
15	73.095238	sign	3721.0213	pen	1.0590163	effects
16	68.980873	top	3573.9697	interest	0.991521	composition
17	64.911892	english	3559.3032	line	0.9701349	planet
18	64.911892	prof	3508.0264	web	0.9024053	pros
19	64.911892	student	3274.7602	point	0.8771075	mistakes
20	60.890449	program	3260.2448	way	0.8452535	audience
21	56.918913	gain	3252.99	self	0.8409857	commerce
22	56.918913	lectures	3238.4862	site	0.8144002	tables
23	52.999902	ease	3231.2372	present	0.8117751	importance
24	52.999902	material	3223.9901	view	0.8061971	window
25	52.999902	star	3202.2605	code	0.7730678	desire
26	49.13632	clot	3115.5187	students	0.7724119	wishing
27	49.13632	content	2964.4163	prof	0.765526	task
28	49.13632	topic	2871.3347	experience	0.7409395	browser
29	45.331414	courses	2857.0463	courses	0.6901589	team
30	45.331414	experience	2857.0463	material	0.6813176	learners
31	45.331414	line	2821.3631	things	0.6279696	waste
32	45.331414	pen	2807.105	instructor	0.6214271	manner
33	45.331414	professor	2778.615	bit	0.6106335	development
34	45.331414	question	2750.1603	cons	0.6099795	someone

Table 9: Overview of salient topics according to drop-out and negative reviews.

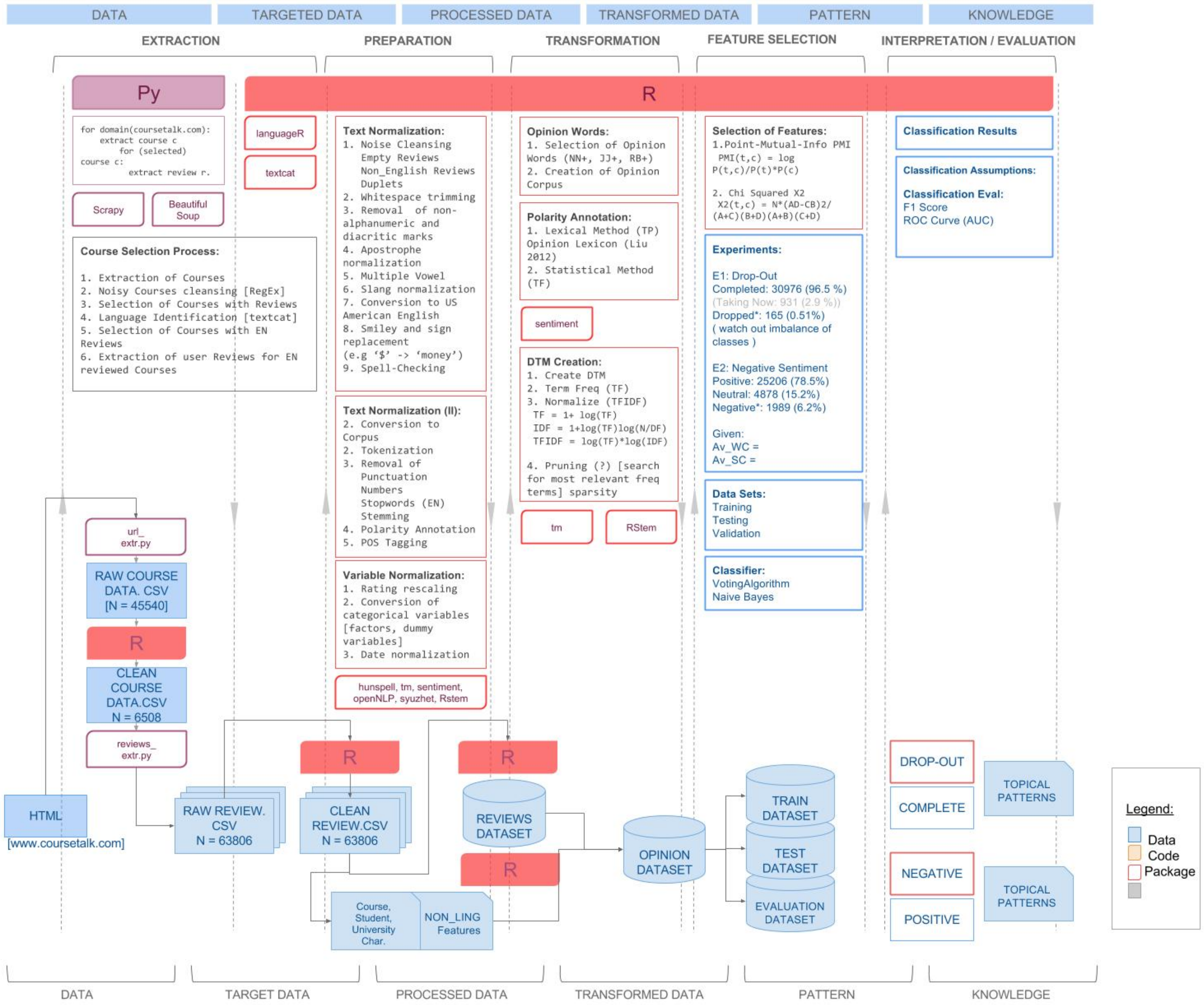

Figure .1: Block diagram representing the KDD process along with the elements involved. Block diagram describing the KDD process along with the elements involved. The input and output are described with the arrows. Source: Author's representation

Eigenständigkeitserklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Datum, Ort

Student