DEGREE PROJECT IN TECHNOLOGY,
FIRST CYCLE, 15 CREDITS
STOCKHOLM, SWEDEN 2018

The Viability of Machine Learning Models Based on Levenstein Distance and Cosine Similarity for Plagiarism Detection in Digital Exams

ELIZABETH ANZÉN

KTH ROYAL INSTITUTE OF TECHNOLOGY
SCHOOL OF INDUSTRIAL ENGINEERING AND MANAGEMENT

The Viability of Machine Learning Models Based on Levenstein Distance and Cosine Similarity for Plagiarism Detection in Digital Exams

Elizabeth Anzén

***Abstract*—This paper investigates the viability of a machine learning model based on similarities in text structure compared to one based on statistical properties in the text to detect cheating in digital examinations. The machine learning model comparing similarity in text structure used Levenstein distance and the one comparing statistical text properties compared cosine distance between word vectors.**

The paper also investigates whether security has been a driving force impacting the industrial dynamics of the digitalization of examinations in Sweden. This is done using the multi-level perspective framework and interviewing users of a digital examination platform.

The results show that the machine learning model based on statistical text properties has a higher accuracy, recall, precision and F-score. Nothing is concluded from this, however, due to discussion of validity of the results from the machine learning model based on the similarities in text structure. The analysis of the industrial dynamics shows that security has been a driving force towards digitalization.

***Index Terms*— Machine Learning, Digital Examinations, DigiExam, Industrial Dynamics, Technological Innovation Systems**

I. INTRODUCTION

Digitalization of examinations is a current trend in the Nordic region. Since January of 2014, all examinations at Syddansk universitet (SDU) have been digital (Nilsen et al., 2014) and Norges teknisk-naturvitenskapelige universitet (NTNU) has set a goal that all examinations should be digital by 2019, (Sindre & Vegendla, 2015). In Sweden Skolverket has announced that the national exams will be digitalized (Skolverket, 2018). One of the leading actors in the Swedish market is the Digital examination platform provider DigiExam. The company has offered a digital examination platform since 2014. The company has enabled this study by providing a large dataset of digital examinations, access to their customers and knowledge about the technical transformation process of examinations.

For digital examination platforms to be a viable alternative to traditional exams they must include countermeasures against cheating. Sindre & Vegendla (2015, p.3) define cheating as "behavior which is against the regulations of the university or of the particular exam, and which may give some candidates an unfair advantage over others.". Sindre & Vegendla (2015) list the following new types of cheating threats that especially occur with digital exams:

1. Impersonation
2. Collaboration
3. Plagiarism
4. Using aids that are not allowed,
5. Time violations
6. Lying to proctors,
7. Bringing the exam out of the classroom.

Whether the frequency of cheating vary in traditional exams and digital exams has been researched with contradictory results. Grijalva, Kerkvliet, Clifford (2006) found no significant difference in the extent of cheating in digital exams compared to traditional exams. Stuber-McEwen, Wisely, Hoggatt (2009) found that cheating occur to a wider extent in traditional exams. Hoggatt et.al (2009) found that the most common way of cheating in an exam regardless of if it is digital is "aiding and abetting".

The problem of detecting cheating can be formalized to the problem of finding similarity between documents.

10-05-18. This work was supported and enabled by the Company DigiExam.
E. Anzén., with KTH, Royal Institute of Technology. Stockholm. Sweden. eliekm@kth.se

(Lukashenko, Graudina, Grundspenkis, 2007). Baba, Nakato, Minami (2016) list two main approaches to finding document similarities. The first one is using a bag of words model which is a statistical approach and the second one is comparing patterns in word occurrences, hereby referred to as structural approach.

II. OUTLINE OF THIS PAPER

The overall aim of this paper is to investigate whether a machine learning model based on clustering and document similarity can be a viable method for detection of plagiarism.

A. Problem Statement Regarding Computer Science

This paper aims to investigate whether cheating in a data set of exams in Swedish are detected more efficiently with a machine learning method built on a statistical approach or a structural approach.

B. Problem Statement Regarding Industrial Management

The industrial management section of this paper will analyze whether development of tools to detect cheating will be a driving force or barrier that impacted the industrial dynamics in the shift towards digital examinations

III. THEORETICAL BACKGROUND

A. Statistical Approaches

Common statistical methods include fingerprinting and term frequency matrixes (TFM). A TFM is a matrix where each column is a document and the rows show frequencies of each word. Common issues with implementation of TFM occur due to high dimensionality and include varied results and high time consumption. Each column in a TFM is a vector with word frequencies. The similarity between the documents in the matrix can be calculated using several metrics. Some common metrics to measure the distance between vectors include Euclidean distance, Manhattan distance and cosine distance. The definition of the Euclidean distance between a vector*, u,* and a vector *v* of *length n* is given in equation (1).

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{n}(u_i - v_i)^2} \qquad (1)$$

The definition of the Manhattan distance between a vector u and a vector v of length n is given in equation (2).

$$\text{Manhattan Distance} = \sum_{i=1}^{n}|(u_i - v_i)^2| \qquad (2)$$

The definition of the cosine distance between a vector v and a vector u is given in equation (3).

$$\text{Cosine Distance} = \frac{u*v}{\|u\|_2\|v\|_2} \qquad (3)$$

Krisnawati & Schulz (2013) conclude statistical approaches based on word frequencies is the appropriate tool when the aim is to reach high precision. They argue that the fact Kong's algorithm won 1th price in the international plagiarism competition proves this.

B. Structural Approaches

Methods using the patterns in word occurrences approach include the Smith Waterman Algorithm and Levenstein distance, (Baba et al, 2016). The Levenstein distance measures how much a string, *a,* needs to be altered to become another string, *b*. The distance is a function of how many characters that needs to be removed, exchanged or added.

Equation (4) states the used definition of the Levenstein distance. The Levenstein Distance between two *strings u* and *v* of length |u| and |v| respectively is denoted $lev_{u,v}(|u|, |v|)$.

$$lev_{u,v}(i, j) = \begin{cases} \max(i,j) \\ \min\begin{cases} lev_{u,v}(i-1, j) +1 \\ lev_{u,v}(i, j-1) +1 \\ lev_{u,v}(i-1, j-1) +1 \end{cases} \end{cases} \qquad (4)$$

C. Clustering

Clustering is a type of unsupervised machine learning where data is sorted automatically. Wu (2012) lists the following five main types of clustering algorithms:

- Prototype based clustering algorithms - These algorithms create a prototype for every cluster and cluster the data points around the prototype.
- Graph-based algorithms - These algorithms see the data set as a graph. The data points are nodes and the distance between them is the weight of the graph lines. A cluster is a closed cycle in the graph.
- Density based cluster algorithms – The general principal on which density based cluster algorithms create clusters is that clusters are areas where data points occur with higher density.
- Hybrid Algorithms – These algorithms use two or more clustering algorithms in combination.
- Algorithm-Independent Methods – The algorithms use the clustering result of basic clustering algorithms instead of the original data.

Agglomerative hierarchical clustering (AHC) can be considered a graph-based method (Wu, 2012). AHC uses a "bottom-up" approach to cluster the data. It starts with each observation and creates clusters by merging

observations until all are a member of a single cluster. (Myatt & Johnson, 2014). The data points can be merged based on several different criterions. One such criterion is Ward's method. It concludes that the distance between two clusters *A* and *B* is how much the sum of squares will increase when merged (Carnegie Mellon University, 2009). These clusters can be illustrated as dendrograms. To measure how well dendrograms persevere the pairwise distances between the original data points. (Carr, Dorthe, Young, Chris, Aster, Richard & Zhang, 1999)

D. URKUND

URKUND is a Swedish company that provides an automated service to deal with plagiarism issues. URKUND compares documents to three main sources, the internet, publisher content and student content. The service returns the similarity of a provided document to other sources as percentage.

The service is supposed to be a complement to human judgement and what percentage is to be considered plagiarism vary between subjects. In this paper the documents that URKUND marked as a source with similarity from the dataset are regarded as cheating.

E. Innovation Systems

An innovation system is defined as "an interrelated structure of institutional and actor based condensations in an economic space" according to Laestadius & Rickne (2016). There are several approaches to innovation system e.g. technical, national and sector. The set of approaches can together be said to form the innovation system framework (Laestadius & Rickne, 2016).

In this paper technological innovation system (TIS) are addressed. Carlson and Stankiewicz (1991) who introduced the term we today refer to as TIS as follows, "Dynamic network of agents interacting in a specific economic/industrial area under a particular institutional infrastructure and involved in the generation, diffusion, and utilization of technology.

F. Industrial Dynamics

Carlsson (2016) states that the study of industrial dynamics has the following main themes:

1. The causes of industrial development and economic growth, including the dynamics and evolution of industries and the role of entrepreneurship.
2. How the boundaries of the firm (degree of vertical and horizontal integration) and the degree of interdependence among firms change over time and what role this interdependence plays in economic growth.
3. Technological change and its institutional framework (particularly in the form of "systems of innovation").
4. The role of public policy in facilitating or obstructing adjustment of the economy to changing circumstances (domestically as well as internationally) at both micro and macro levels.
5. The nature of economic activity in the firm and its connection to the dynamics of supply and therefore economic growth, particularly the role of knowledge (competence).

Arvidsson (2016) states that research on dynamics in innovation systems has shown that there are factors conserving a system as well as factors that are changing the system. Dynamic processes are shaped by the interaction between both of these.

Several theoretical models and frameworks can be used to analyze the industrial dynamics behind industry transformation. Two commonly used frameworks are the Multi-Level Perspective (MLP) and the Large Technical System (LTS). LTS is suitable to study large physically connected infrastructure systems. MLP is appropriate when studying societal and technological transitions (Blomkvist & Johansson, 2016). The MLP approach was developed by Geels (2002). The MLP framework separates the studied area or sector into the following three system levels:

1) Niches are secure environments where innovation can mature and be tested. An example of a niche is a military research project. (Blomkvist & Johansson, 2016). A problem with niche innovations is that they can be misaligned with existing technical infrastructure and regulations (Schot & Geels 2008).
2) The second system level is Socio-technical regimes which is considered the meso level. The term "technological regime" was coined by Nelson and Winter (Blomkvist & Johansson, 2016). Geels expanded the definition to the following (2004: 900) "... define ST-systems in a somewhat abstract, functional sense as the linkages between elements necessary to fulfil societal functions (e.g. transport, communication, nutrition). As technology is a crucial element in modern societies to fulfil those functions, it makes sense to distinguish the production, distribution and use of technologies as sub-functions. To fulfil these sub-functions, the necessary elements can be characterized as resources. ST-systems thus consist of artefacts, knowledge, capital, labour, cultural meaning". Socio-technical regimes are comprised of three dimensions according to Geels (2004):
    - Network of actors and social groups.
    - Formal, normative and cognitive rules
    - Physical and technical elements.

An example of a technological regime provided by Nelson and Winter was the airplane Douglas DC-3. It had completely revolutionary features such as all metal skin and a low wing. These features influenced all new airplane design (Blomkvist & Johansson, 2016).

3) The third level is the macro level and is called sociotechnical landscapes. The sociotechnical landscape is similar to what system theorists refer to as the system environment. It contains the intuitional and market aspects. (Blomkvist & Johansson, 2016).

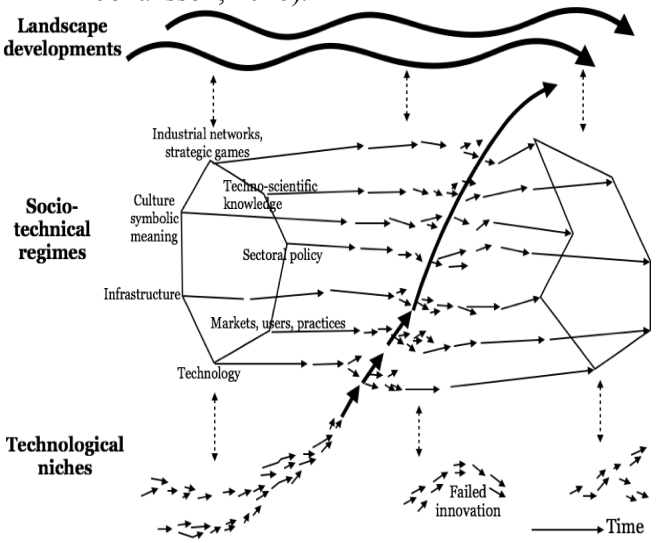

Fig. 1. Illustration of the multi-level perspective, adapted from Geels (2002)

IV. METHOD

A. General Description of Method

Two separate machine learning models were created. One was based on a structural approach and another one based on a statistical approach. For both machine learning models, data was prepared and clustered in accordance with the description in D, E.

For the structural approach, a model was based on the average Levenstein distance in each cluster.

For the statistical approach, a model was built based on the average cosine distance in each cluster. Equation (2) shows the definition of the cosine distance used.

Both models assigned the test data points to clusters whereby they were classified as cheating or not based on whether the similarity was greater than the average for the assigned cluster.

B. Evaluation

The machine learning models were evaluated with the measurements accuracy, recall, precision, and F-score. The definition of the measurements is specified in equations (5), (6), (7) and (8). In the equations, the term "true positives" (TP) refers to exams that both Urkund and the machine learning model classify as cheating. The term "true negatives" (TN) refers to exams that both Urkund and the machine learning model classify as honestly produced. The term "false positives" (FP) refers to exams that Urkund classifies as honestly produced but the model classifies as cheating. The term "false negatives" (FN) refers to exams that Urkund classifies as cheating but the model as honestly produced.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{F-Score} = \frac{2}{\frac{1}{\text{Recall}} + \frac{1}{\text{Precision}}} \quad (8)$$

C. Dataset

The machine learning model that this research was based upon used a dataset provided by the company Digixam. The data set was anonymized. The dataset was comprised of 11 974 exams in Json format done by Swedish students. The exams were at least 20 words or longer. Out of these, 9759 exams were used as a training set and 2215 were used as a test set. To measure the quality of the data for clustering, cophenetic correlation was calculated. The calculation was done using the module cophenet from scipy.cluster.hierarchy.

D. Preparation of Data for the Clustering and the Statistical Approach

The preparation of the dataset was done through removal of HTML-tags, removal of stop words (see appendix) and stemming and parsing of the texts. The stemming was done using the nltk module Snowball stemmer. The term "stop words" refer to words that bear little or no semantic meaning, e.g. conjunctions.

Vectorization of the text was conducted using TfidfVectorizer from the module Feature_extraction.text in sklearn.

The input parameters that have been used are max_df = 0.9, min_df=0.1, max_feature=500, lowercase = False, stop_words = None, use_idf = False, Tokenizer = None, ngram_range(1,3). The parameter max_df=0.9 means that words that occur in 90% of the documents or more are removed and the parameter min_df =0.1 means that words that occur in less than 10% of the documents are removed from the vector. The words are already lowercased, tokenized and Swedish stop words which is why the parameters lowercase is False and the stop_word and tokenizer parameters are None. The parameter use_idf is the option to enable inverse-document frequency reweighting. This is not deployed since the documentation on how the inverse document frequency reweighting is done is limited.

E. Clustering

To make the model computationally viable and to avoid unnecessary comparisons, clustering was deployed. The chosen method is hierarchal Ward clustering.

To calculate the linkage method, the module ward from scipy.cluster.hierarchy was used. The input parameter was a distance matrix with the cosine distance between the vectors in the representation returned from TfidfVectorizer. The cosine similarity was calculated using the cosine_similiarity module from sklearn.metrics.pairwise.

The clusters were illustrated in a dendrogram using the module dendrogram from scipy.cluster.hierarchy, (figure 2). The maximum distance between clusters was estimated from the dendrogram.

The clusters were extracted using the module fcluster from scipy.cluster.hierarchy. The module was used with the input parameter *Z* as the linkage matrix returned by the ward module. The *t* parameter was set to the maximum distance which is 140. The parameter *criterion* will be set to maxclust. This allowed the method to find a threshold *r*, so that the cophenetic distance between any two original observations in the same flat cluster wass no more than *r*.

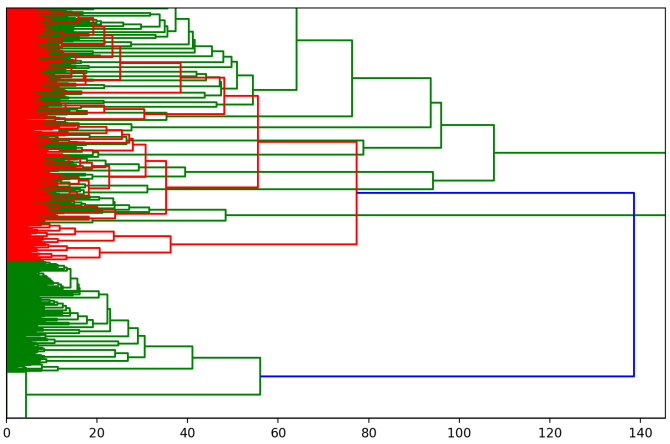

Fig. 2. Dendrogram performed on the training dataset.

F. Preparation of the Dataset for the Structural Approach

For the structural approach each exam answer was parsed into a list of sentences. Stop words were not removed and stemming was not performed.

G. Structural Approach

The machine learning model created based on the structural approach was trained in accordance with the following steps:

1. Prepare the dataset for clustering in accordance with the description in D.
2. Cluster the data using hierarchal clustering in accordance with the description in E.
3. Prepare the dataset for comparisons of Levenstein distances in accordance with the description in F.
4. Calculate the Levenstein distance for each sentence compared to other sentences in the same cluster that have more than 20 words or items of punctuation in common.
5. Calculate the average Levenstein distance for each cluster.

The machine learning model created based on the structural approach was tested in accordance with the following steps:

6. Assign the data points in the test dataset to clusters using a K-nearest neighbor classifier with the module KNeighborsClassifier from sklearn.neighbors.
7. For each exam in the test data set, calculate the Levenstein distance between each sentence in the exam and all other sentences that have 20 or more words or items of punctuation in common in the same cluster.
8. If the Levenstein distance for the sentence is greater than the average for the cluster it will be saved, and a cheating warning will be issued.

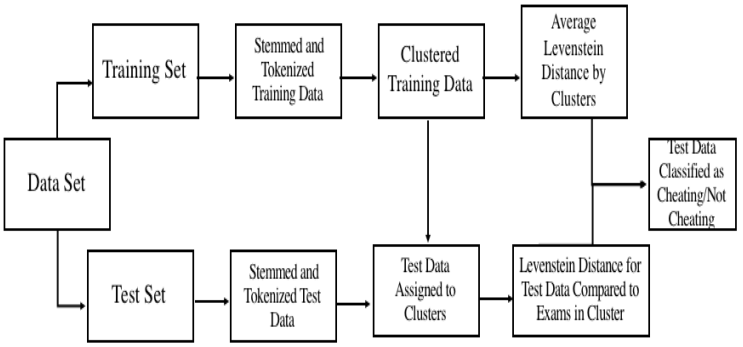

Fig 3. Illustration of the calculations and modifications done in the structural model.

H. Statistical Approach

The machine learning model created based on the statistical approach was trained in accordance with the following steps:

1. Prepare the data set for clustering in accordance with the description in D.
2. Cluster the data using hierarchal clustering in accordance with the description in E.
3. Calculate the average cosine distance in each cluster.

The machine learning model created based on the statistical approach was tested in accordance with the following steps:

4. Assign the data points in the test dataset to clusters using a K-nearest neighbor classifier with the module KNeighborsClassifier from sklearn.neighbors.
5. For each exam in the test data set, calculate the cosine distance between the exam and all other exams in the cluster it was assigned to.
6. Compare the calculated cosine distance for the exam to the average cosine distance for the cluster. If the calculated cosine distance is greater than average, a cheating warning will be issued.

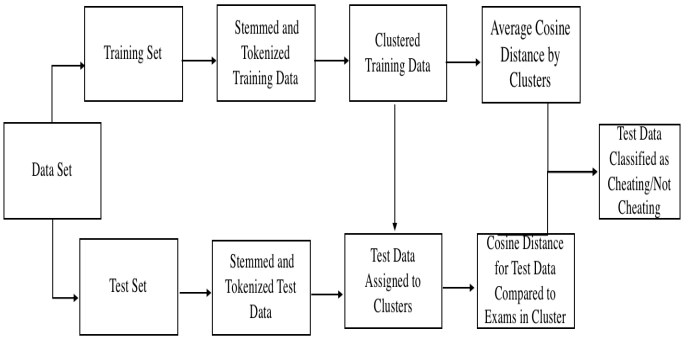

Fig 4. Illustration of the calculations and modifications done in the structural model.

I. Method to Investigate Industrial Management Problem Statement

To investigate the industrial management problem statement, the method chosen was to conduct three interviews with customers of the company DigiExam. The interviewees are teachers who are familiar with DigiExam's platform. One interviewee had extra responsibility for spreading IT-knowledge within the school organization. The interviews where held in Swedish and the questions can be viewed in the appendix. The interviews have been transcribed.

Aside from the interviews, a review of literature on digital examinations and security was done. The literature reviewed was Björklund and Wenestam (1999), Sindre and Vegendla (2015) and Heintz (2017).

The results from the interviews and the literature review was the basis of an analysis of the industrial dynamics and whether cheating detection had been a driving force. The analysis was done using the multi-level perspective framework.

V. Results

A. Quality of Dataset

The cophenetic correlation for the training data set and the test data set was calculated. The result was calculated as 0.45 for the training data set and 0.51 for the test data set. Table 1 shows the cophenetic correlation results for the data set in general.

TABLE I
COPHENETIC CORRELATION FOR DATA SET

Data set	Cophenetic Correlation
Training Data	0.45
Testing Data	0.51

B. Results from URKUND

Out of the test data set that consist of 2215 documents, URKUND marked 176 of them as a source with some degree of similarity to another document in the data set.

C. Results from Statistical Approach

Out of the 2215 documents in the test data set the statistical approach marked 1377 of them as cheating. Table 2 shows the number of TP, FP, TN and FN.

The results for the evaluation metrics accuracy, recall, precision and F-score from the statistical approach are shown in table 3. The accuracy for the test dataset was 40.6 %. The precision for the test data set was 8.5%. The recall for the test data set was 67.0%. The F-score was thereby calculated to 15.2%.

TABLE 2
RESULTS FROM STATISTICAL APPROACH

Category	Approach	Number
True Positives	Statistical	118
True Negatives	Statistical	780
False Negatives	Statistical	58
False Positives	Statistical	1259

TABLE 3
RESULTS FROM STATISTICAL APPROACH

Metric	Approach	Value
Accuracy	Statistical	40.6 %
Recall	Statistical	67.0 %
Precision	Statistical	8.5 %
F-score	Statistical	15.2 %

D. Results from Structural Approach

Out of the 2215 documents in the test data set the structural approach marked 7 of them as cheating. Table 4 shows the number of TP, FP, TN and FN.

TABLE 4
RESULTS FROM STRUCTURAL APPROACH

Category	Approach	Number
True Positives	Statistical	1
True Negatives	Statistical	2033
False Negatives	Statistical	175
False Positives	Statistical	6

The results for the evaluation metrics accuracy, recall, precision and F-score from the structural approach are shown in table 5. The accuracy for the test dataset was 91.8 %. The precision for the test data set was 14.2%. The recall for the test data set was 0%. The F-score was thereby calculated to 0.01%.

TABLE 5
RESULTS FROM STRUCTURAL APPROACH

Measurement	Approach	Value
Accuracy	Structural	91.8%
Recall	Structural	0.06%
Precision	Structural	14.2%
F-score	Structural	0.01%

E. Summation of Interviews from the Industrial Management Section

Three interviews have been conducted. Summations of the interviews are listed below.

The first interviewee had a formal role as a teacher in and was responsible for sharing IT-knowledge. The interviewee taught history and social science. The interviewee was a teacher at a school in the municipality Nacka.

1) *Below follows a summation of the first interview:*

At the interviewee's schools there are no central rules in regard to how other examinations than the national exams are held. The interviewee perceived it as a professional freedom. Examination rules had not changed since the school started using digital examinations.

The school has used digital examinations since 2014 the transition was initiated from the political governance in the municipality. The interviewee states that one reason for why the school decided to use the digital examination platform DigiExam rather than its competitors was that it was perceived as a safer option. Some features of DigiExam that were perceived as safe were the ability to lock down the user's computer and the ability to use the platform offline. The interviewee cannot state whether he perceives digital examinations as more or less safe than traditional examinations. The reasoning presented for this is that the interviewee perceives them as safer but have colleagues who do not.

There is a policy at the school regarding what teachers are supposed to do if they discover cheating among students, but it has not changed since the school started with digital examinations. The interviewee says that the policy had already been adapted to the digital era due to the rising use of smart phones among students.

The interviewee does not know whether the frequency of cheating has increased or decreased since the school chose to use digital examinations.

The interviewee was asked to examine the means of cheating that are likely to increase with digitalization of examination stated by Sindre and Vegendla (2015). The interviewee had not encountered all the means listed.

When asked to rank them by frequency the interviewee listed them in the following order:

1. Collaboration
2. Plagiarism
3. Use of aids that are not allowed
4. Imitation
5. Violating time limits
6. Bringing questions from exam
7. Lying to proctors.

When asked to rank the listed means of cheating by which a protection feature would add most value the interviewee ranked them in the same order.

2) *Below follows a summation of the second interview.*

The second interviewee was a teacher in the region Västerås and taught religion, psychology and religion. The interviewee had been a teacher for twelve years and has worked with digital examination for two years.

There were currently no central rules at the school were the interviewee worked and there have never been any. Some teachers at the school allows students to use cellphones for example.

Two main reasons for transitioning to digital examination were stated:

- Digital examinations are easier to read and make the examination process more efficient.
- The second stated reason was to make sure students do not cheat in the same way. However, the interviewee states that they still can use aids that are not allowed.

When the transition to digital examinations was made, the digital examinations were perceived as safer. However, students have been observed using the platform in a strange manner lately. This behavior had made the interviewee uncertain on whether digital examinations are safer or not.

The interviewee described that examinations are now held in a different way since the school started using digital examinations. Essay questions that students get to work on at several occasions are created. The interviewee stated that this is done to avoid parents and siblings doing the assignments and students plagiarizing.

The interviewee had an overall perception that the students cheat more now than prior but less cheating is detected.

The interviewee was asked to examine the means of cheating that are likely to increase with digitalization of examination stated by Sindre and Vegendla (2015). The interviewee had not encountered all the means listed.

When asked to rank them by frequency the interviewee listed them in the following order:

1. Use of aids that are not allowed
2. Lying to proctors
3. Collaboration
4. Bringing questions from the exam
5. Plagiarism
6. Violating time limits
7. Impersonation

When asked to rank the listed means of cheating by which a protection feature would add most value the interviewee ranked them in the same order.

3) *Below follows a summation of the third interview:*

The third interviewee was a teacher in the Västerås area and mainly taught English and athletic psychology. The interviewee had worked with digital examination for two years. There is currently a central rule that cellphones are supposed to be collected prior to an examination but the rule was in place prior to digitalization of examinations. Two main reasons for transitioning to digital examinations was provided,

- Firstly, students mostly spend their time writing on a computer not by hand. Thereby unnecessary transition should not be made,
- The DigiExam digital examination platform provided lock down functionality.

The teacher perceives digital examinations as significantly safer than traditional examinations.

The teacher also perceives that cheating has decreased since the school transitioned to digital examinations. There is a policy with steps to be taken at the school if a teacher detects cheating. The principal and parents should be informed.

The interviewee was asked to examine the means of cheating that are likely to increase with digitalization of examination stated by Sindre and Vegendla (2015). The interviewee had not encountered all the means listed.

When asked to rank them by frequency the interviewee listed them in the following order:

1. Bringing exams from the classroom
2. Use of aids that are not allowed
3. Collaboration
4. Impersonation
5. Plagiarism
6. Lying to proctors
7. Breaking time limits

When asked to rank the listed means of cheating by which a protection feature would add most value the interviewee ranked them in the same order.

1. Use of aids that are not allowed
2. Collaboration
3. Impersonation
4. Plagiarism
5. Lying to proctors
6. Imitation
7. Breaking time limits

VI. Discussion

A. Choice of Similarity Level Between Sentences in the Model Based on the Structural Approach

The high accuracy from the structural approach compared to the statistical approach does not provide support for the conclusion that the structural approach is more viable. This due to the poor results on precision and recall. The validity of the results from the structural approach are discussed in B.

Possible reasons for the poor precision and accuracy result from the statistical approach are discussed in E. The relative high level of recall for the statistical approach does however indicated that it has the potential to become a viable method with altercations.

B. Validity of results from structural approach

To make the machine learning model more viable with respect to time complexity, the choice was made to only compare sentences in a cluster that had 20 words and items of punctuation in common. This may have skewed the results towards a higher cluster average. In turn, this may have resulted in fewer documents detected and a higher number of FN. An improvement of the model could have been achieved by conducting the experiment with a lower level of similarity between compared sentences. Possibly, this could reduce the number of FN.

C. Choice of Similarity Measurements

The chosen measurement in the structural approach in this paper was the Levenstein distance. There are several possible alternatives e.g. the Smith Waterman algorithm or the normalized Levenstein distance proposed by Li & Liu (2007). Whether the results would have been similar for the structural approach if it had been based upon one of these measurements is a possible subject to further study.

The selected measurement for the statistical approach was the cosine distance between the word vectors in a TFM. The cosine distance was selected rather than the Euclidean distance or the Manhattan distance since it can be calculated for vectors of different size.

D. Choice of Clustering Method

One factor that may have affected the results is the choice of clustering method. One aspect that was considered in the choice of clustering method is if it required a desired number of clusters. A clustering method that require a desired number of cluster would have resulted in that the quality of the model was dependent on the knowledge of the dataset. There are several possible clustering methods that satisfy that condition aside from hierarchal ward clustering for e.g. DBSCAN. Whether the results would have been improved by using another clustering method is a possible subject to future study.

In order to generalize the results, it was important to address whether the choice of clustering method had affected the results. This research does not provide any evidence on the viability of a machine learning model based on the structural approach compared to a machine learning model based on the statistical approach independent off clustering method.

E. Evaluation

Another factor that may have impacted the results is that both of the models issue a cheating warning when the similarity is greater than the average in the cluster the exam is assigned to. Greater similarity between two documents than the average in the cluster does not necessarily mean that the authors of the documents have cheated. This may have affected the results to have a greater number of false positives then if a higher level of similarity to issue a warning of cheating had been chosen. Thereby, the low precision level from the statistical approach does not contradict previous work by Krisnawati & Schulz (2013) that states that statistical approaches based on word frequencies are the appropriate method when high precision is the goal.

The desired sensitivity of the model is dependent on the individual usage case. If the model will be complemented with human judgement a higher level of sensitivity is presumed to be preferred.

F. The Use of URKUND

The evaluation consists of comparing how well the suggested methods detect document similarity, compared to the method for detecting document similarity used by URKUND. To investigate how well this measures document similarity, other services that compare document similarity, like Turnitin, could be used. If several state-of-the-art several services that find document similarity mark the same documents, it would validate the results.

Other possible and more suitable methods to evaluate the experiment include using a data set where the exam has been marked manually as cheating by a teacher, or a data set that consists of determined cheating cases. Both these methods would have required an extensive manual work or a large data set that was not available for this research.

Another aspect of the evaluation is at what level of similarity by URKUND the documents are marked as cheating. In this paper everything above zero is used. However, this may have marked documents as cheating that for example cite the same person. It is, however, not possible to eliminate this risk, as quotes and sources can be of any size. For services aimed to be used as a complement to human judgement, a lower threshold can therefore be assumed to be desirable.

G. Run Time

An aspect of whether you can consider the model viable or not is runtime. The runtime for the machine learning model based on the structural approach was approximately 22 hours which include reading data, parsing data, clustering, calculating the model and testing the model. The runtime for the machine learning model based on the statistical approach was approximately four hours. This supports the conclusion that the statistical approach can be considered more viable.

VII. Industrial Dynamics in Digitalization of Examinations in Sweden

In this section the industrial dynamics of the digitalization of examinations are analyzed. The analysis is comprised of a list of relevant actors, the factors impacting the industrial dynamics identified in the interviews and literature review and their relation to the MLP-framework.

A. Relevant Actors

The main actors relevant to analyze the industrial dynamics of digital examinations are listed below:

- Teachers – The teachers main role is to create and oversee digital examinations.
- Digital examination platform providers – Companies that provide digital examination platforms. Some major actors on the Swedish market include DigiExam and Dugga.
- Students – Users of digital examination platforms.
- Schools – Institution to spread knowledge where examinations are held. The demand for a digital examination platform are determined on a school by school basis.
- Municipalities – In Sweden the municipalities are responsible for running and funding the schools. There are instances where they have purchased license for a digital examination platform for all public schools in a municipality.

- Government Agencies – Oversight agencies like Skolverket and Skolinspektionen.

B. Identified Factors Relation to the MLP-framework

The identified factors influencing the industrial dynamics related to digitalization of examination have been identified and sorted as landscape developments, socio-technical regime and technological niches. The landscape developments are changes in the entire school system. The socio-technical regime is related to public policy and common practices. The technical niches are mainly educational tech companies that develop new parts of the system.

C. Identified Factors

From the literature review it was concluded that an important factor driving towards digitalization of examination is organizational or political goals. This conclusion was supported by the result from the first interview. These are both factors on the socio-technical regime level in MLP-framework and both in the dimension formal, cognitive and normative rules. Organizational goals impact the normative rules and the overall idea of what an examination can be among educators. Political goals impact the formal rules and create pressure on organizations to change.

Another important factor is the emergence of digital examination platforms. When the development of digital examination platforms started it was on the niche level. This combined with pressure on the landscape level from the overall trend of digitalization created opportunity for diffusion of the niche innovation digital examination platforms. Since these processes were aligned it allowed a breakthrough of these innovations.

The first interview indicated that a factor that may act as a barrier is the level of competence among staff in IT. It will be addressed as a factor but to validate them as relevant more research should be conducted. This could act as a barrier on the regime level if the transformation is reject by a key actor like teachers.

The third factor driving the change is process improvement of exams. Sindre and Vegendla (2015) lists the following three improvements of the examination process that stems from digitalization of exams:

- Development of the examinations.
- Increased efficiency of the examination process.
- Simplified assessment of exams.

All interviews indicate that the digitalization of examinations allows for more development and improvement of examinations and the examinations process. Process improvement creates pressure manifests as pressure on the socio-technical regime. It impacts the dimension network of actors and social groups.

Three main aspects impacting whether security is a driving force, or a barrier has been identified.

The first aspect, is whether digital examinations are perceived as more or less safe by the teachers. In the conducted interviews contradictory views on the matter were expressed. This does not provide any evidence on whether security is a factor driving towards change or acting as a barrier. This is a factor on the socio-technical regime level,

The second aspect is the frequency of cheating in digital examinations compared to cheating in traditional exams. The research in the area is contradictory as discussed in the introduction. The results from the first interviews are contradictory on this subject. Results from Hoggatt et.al (2009) indicate that the most common way to cheat regardless of examination type is collaboration or aiding and abetting. This was supported by results from the first qualitative interview where collaboration was listed as the most frequent mean of cheating. Since there are increased possibilities to detect collaboration with digital examinations (by comparing document similarity) this would indicate that security concerns are a factor driving towards change. These factors affect the socio-technical regime level and mainly the dimension actors and social groups.

The third factor is new means of cheating that occur with digital examinations. Research by Heintz (2017) conclude that digital examinations where the student brings their own device increase the risk for advanced cheating attempts such as DDOS-attacks and code injection attempts. The current understanding of these risks in Swedish schools are presumed to be low. This supports the conclusion that it does not act as a barrier. The awareness of these risk may increase in future which could create pressure on the landscape level. The landscape pressure could result in changes on the regime level in the dimension formal rules. This could in future result in the breakthrough of new niche innovations regarding increased security. From this I conclude that safety has been a force driving towards digitalization of examinations but will in future act as a barrier.

In table 6 the main factors are listed and categorized as a dimension of the socio-technical regime.

TABLE 6
FACTORS DRIVING TOWARDS DIGITALIZATION OF EXAMINATIONS

Factor	Barrier or Driving change	Level in MLP-framework
Organizational Goals	Driving Change	Socio-technical regime – Normative rules dimension
Political Goals	Driving Change	Socio-technical regime –Formal rules
Process improvement	Driving change	Socio-technical regime - Network of actors and social groups
Emergence of digital examination platform	Driving change	Niche
Security	Driving change	Socio technical regime
Level of IT-competence	Barrier	Socio-technical regime - Network of actors and social group.

D. Discussion on Method Regarding Industrial Management Section

To validate the results from the industrial management section of the report more interviews could have been conducted. Another possible improvement could have been to select interviewees that all had formal roles or responsibility related to IT and not only experience from using the tool for digital examinations.

One could argue that a quantitative study to complement my qualitative research would have improved the method. However, the area of frequency of cheating is well researched as described in the introduction. An additional survey is unlikely to have resulted in a more nuanced picture than the one that gained from reading the current research.

A subject for further study is to investigate whether perceived cheating frequency is dependent on the academic level or grade point average of the students in a school.

VIII. CONCLUSION

A. Conclusion from Machine Learning Section of Paper

No conclusions on whether a machine learning model based on a statistical approach is more effective than one based on structural approach can be drawn. This due to the issues of validity of the results from the structural discussed. However, it can be concluded that the statistical approach is more viable with respect to time complexity.

B. Conclusion from Industrial Management Section of Paper

From the analysis done of the qualitative interviews it was concluded that security has been a driving factor impacting the industrial dynamics in digitalization of examinations. However, it was also concluded that this is likely to change in future creating new pressure and enabling new security innovations.

ACKNOWLEDGMENT

I would like to thank John Reuterswärd, Carl Nardini and Robin Andersson at DigiExam. Without their help with resources and guidance this would work could not have been done.

I would also like to thank Bo Karlsson and Olov Engwall at KTH for their guidance and help.

REFERENCES AND FOOTNOTES

[1] Arvidsson, N. The Route towards a Cashless Society in Sweden. 2016. *A Dynamic Mind*, KTH.

[2] Baba. K, Nakatoh. T, Minami. T, ,Plagiarism detection using document similarity based on distributed representation, *Procedia Computer Science*, Volume 111, 2017, Pages 382-387, ISSN 1877-0509.

[3] Björklund, M. and C.-G. Wenestam, Academic cheating: frequency, methods, and causes, in European Conference on Educational Research. 1999: Lahti, Finland.

[4] Blomkvist, P. Johansson, P. Systems thinking in Industrial dynamics. 2016. *A Dynamic Mind*, KTH.

[5] Carlsson,B. (2016) Industrial Dynamics: A Review of the Literature 1990–2009, Industry and Innovation, 23:1, 1-61

[6] Carlsson, B. Stankiewicz, R. (1991). On the nature, function and composition of technological systems. *Journal of Evolutionary Economics, 1*(2), 93-117.

[7] Carr, Dorthe B., Young, Chris J., Aster, Richard C., & Zhang, Xioabing. (1999). *Cluster Analysis for CTBT Seismic Event Monitoring*. United States.

[8] Digitalisering av nationella prov: Skolverket. (2018, February 22). Retrieved from http://www.nbcmiami.com/news/local/Teen-Posed-as- https://www.skolverket.se/bedomning/nation

ella-prov/fragor-och-svar/digitalisering-av-nationella-prov-1.265681#utrustning

[9] Distances between Clustering, Hierarical Clsutering : Carnegie Mellon University. (2009, September 14). Retrieved from http://www.stat.cmu.edu/~cshalizi/350/lectures/08/lecture-08.pdf

[10] Geels, F.W., 2002. Technological transitions as evolutionary reconfiguration processes: a multi-level perspective and a case-study. *Research Policy*, 31(8-9), pp.1257–1274.

[11] Geels, F.W., 2004. From sectoral systems of innovation to socio-technical systems. *Research Policy*, 33(6-7), pp. 900.

[12] Geels, F.W. & Schot, J., 2007. Typology of sociotechnical transition pathways. *Research Policy*, 36(3), pp.399–417.

[13] Grijalva. T, Kerkvliet. J, and Clifford. N. Academic Honesty and Online Courses. *Department of Economics, Weber State University,* Ogden, In: College Student Journal (2006).

[14] Heintz, A. Cheating at Digtial Exams – Vulnerabilities and Countermeasures. 2017. Department of Computer Science. Norwegian University of Science and Technology.

[15] Krisnawati. D, and Schulz K. 2013. Plagiarism Detection for Indonesian Texts. In *Proceedings of International Conference on Information Integration and Web-based Applications & Services (IIWAS '13)*. ACM, New York, NY, USA, , Pages 595 , 5 pages.

[16] Laestadius, S. Rickne, A. 2016. A Critical view of the innovation systems approach*, A Dynamic Mind*, KTH 2016

[17] Lanier, M. (2007) Academic Integrity and Distance Learning, *Journal of Criminal Justice Education*, 17:2, 244-261.

[18] Li, Y, Liu B, "A Normalized Levenshtein Distance Metric," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 6, pp. 1091-1095, June 2007.

[19] Lukashenko. R, Graudina. V, and Grundspenkis. T. Computer-based plagiarism detection methods and tools: An overview. In *Proceedings of the 2007 International Conference on Computer Systems and Technologies,* pages 1-6. ACM, 2007.

[20] Myatt, G., & Johnson, W. 2014. Making sense of data I : A practical guide to exploratory data analysis and data mining (2nd ed.).

[21] Nielsen, K. G., Petersen, L., Wallstedt, B., Basse, P., Hansen, P. S., Hansen, S. S., & Majlund Sørensen, D. (2014). ”Evaluation of Digital Assessment”. In *EUNIS conference*, 2014 European University Association.

[22] Sindre, G. and Vegendla, A. 2015b. E-Exams Versus Paper Exams: A Com- parative Analysis of Cheating-Related Security Threats and Countermeasures. In *Norwegian Information Security Conference (NISK)*.

[23] Stuber-McEwen. D, Wiseley. P, and Hoggatt. S. “Point, Click, and Cheat: Frequency and Type of Academic Dishonesty in the Virtual Class- room”. In: Online Journal of Distance Learning Administration 12.3 (2009), pp. 1–9. ISSN: 15563847

[24] Wu J. (2012) Cluster Analysis and K-means Clustering: An Introduction. In: Advances in K-means Clustering. Springer Theses (Recognizing Outstanding Ph.D. Research). Springer, Berlin, Heidelberg

E. Anzén from Stockholm. Currently pursuing a B.S. degree from the Royal Institute of Technology in Stockholm within Industrial Engineering and Management.

APPENDIX

A. Questions Asked at the Interviews or the Industrial Management Section of the Report

Om intervjupersonen:

- Vad har du för roll i skolan?
- Jobbade du här när ni gick över från traditionella till digitala prov?

Bakgrund till varför ni bestämde er för att börja använda digitala prov.

- När införde ni digitala prov?
- I vilken omfattning använder ni digitala prov?
- Vilka var de viktigaste faktorerna som fick er att gå över från digitala prov till traditionella prov?
- Hur resonerade ni kring säkerhet vid provtillfället inför övergången? Var det några specifika aspekter av säkerhet som ni upplevde som extra viktiga?
- Uppfattade ni digitala prov som mer eller mindre säkra än traditionella prov?

Provförfarandet innan digitala prov:

Hur gick provförfarandet till före ni införde digitala prov?

Använder de bring your own device?

- Har förfarandet förändrats någonting sedan ni införde digitala prov?
- Fanns det centrala regler för hur lärarens skulle gå till väga eller har ni infört det?
- Hur har ni tänkt när ni bestämt dem?

har ni läst någon forskning?

har ni utgått från tidigare erfarenhet?

- Hur hanterade ni om ni upptäckte att någon fuskade? Fanns det en policy exempelvis?
- Har denna i så fall behövt uppdateras sedan införandet i så fall?

Fusktekniker:

- Upplever ni att omfattningen av fusk har förändrats sedan ni införde digitala prov? Ökat/minskat?
- Vilka är de vanligaste typen av fusk som ni upptäcker vid provtillfällen? (någon annan gör ett arbete, de använder ett otillåtet hjälpmedel, de tittar på varandra etc)

Det finns en del forskning om vilka de största riskerna för fusk är när man är använder digitala prov. De har identifierat följande huvudrisker,

- *Impersonation – imitation*
- *Collaboration – samarbete*
- *Plagiarism – plagiat*
- *Using aids that are not allowed – använda otillåtna hjälpmedel*
- *Time violations - brott mot tidsgränser*
- *Lying to proctors – ljuga för provvakter om* exempelvis tekniska fel
- Bringing the exam out of the classroom – Att kopiera provfrågor och ta med dem
- Har ni detekterat alla dessa typer av fusk?
- Skulle du kunna ranka dem utifrån vilken som ni upplever som vanligaste efter att ni gått över till digitala prov.
- Skulle du kunna ranka dem utifrån vilken av dem som ni skulle värdera skydd mot högst.

B. Stop words

'aderton', 'adertonde', 'adjö', 'aldrig', 'all', 'alla', 'allas', 'allt', 'alltid', 'alltså', 'andra', 'andras', 'annan', 'annat', 'artonde', 'artonn', 'att', 'av', 'bakom', 'bara', 'behöva','behövas', 'behövde', 'behövt', 'beslut', 'beslutat', 'beslutit', 'bland', 'blev', 'bli', 'blir', 'blivit', 'borde', 'bort', 'borta', 'bra','bäst', 'bättre', 'båda', 'bådas','både','dag', 'dagar', 'dagarna', 'dagen', 'de', 'del', 'delen', 'dem', 'den', 'denna', 'deras', 'dess', 'dessa', 'det', 'detta', 'dig', 'din','dina', 'dit', 'ditt', 'dock',
'dom', 'du', 'där', 'därför', 'då', 'efter', 'eftersom', 'elfte', 'eller', 'elva', 'en', 'enkel', 'enkelt', 'enkla', 'enligt', 'er', 'era' 'ert', 'ett', 'ettusen', 'fall', 'fanns', 'fast', 'fem', 'femte',' femtio', 'femtionde',’femton', 'femtonde', 'fick','fin', 'finnas', 'finns', 'fjorton', 'fjortonde', 'fjärde', 'fler', 'flera', ‘flesta', 'fram','framför', 'från', 'fyra','fyrtio', 'fyrtionde', få', får', fått', följande', för', ’före', 'förlåt', 'förra','första', 'ge', 'genast',' genom', 'ger', 'gick', 'gjorde', ‘gjort', 'god', 'goda', 'godare', godast', 'gott', 'gälla', 'gäller', 'gällt', 'gärna', 'gå', 'gång', 'går', 'gått', 'gör', 'göra', 'ha', 'hade', 'haft', 'han', 'hans', 'har', 'hela', 'heller', 'hellre', 'helst', 'helt', 'henne', 'hennes', 'heter', 'hit', 'hjälp', 'hon', 'honom', 'hundra', 'hundraen', 'hundraett', 'hur', 'här', ‘hög', 'höger', 'högre', 'högst', 'i', ibland', 'idag', 'igen', 'igår', 'imorgon', 'in', 'inför', 'inga', 'ingen', 'ingenting', inget', 'innan', 'inne', 'inom', 'inte', 'inuti', 'ja', 'jag', 'jämfört, kan', 'kanske', 'knappast', 'kolla', 'kom', 'komma', 'kommer', 'kommit', 'kr', kunde', 'kunna', 'kunnat', 'kvar', 'kör', ‘legat', 'ligga', 'ligger', 'lika', 'likställd', 'likställda', 'lilla', 'lite', 'liten', 'litet', 'lägga', 'länge', 'längre', 'längst', 'lätt', 'lättare', 'lättast', 'långsam', 'långsammare', 'långsammast', 'långsamt', 'långt', 'man', 'med', 'mellan', 'men', 'menar', 'mer', 'mera', 'mest', 'mig', 'min', 'mina', 'mindre', 'minst', 'mitt', 'mittemot', 'mot', 'mycket', 'många', 'måste', 'möjlig', 'möjligen', möjligt', 'möjligtvis', 'ned', 'nederst', 'nedersta', 'nedre', 'nej', 'ner', 'ni', 'nio', 'nionde', 'nittio', 'nittionde', 'nitton', 'nittonde', 'nog', 'noll', 'nr', 'nu', 'nummer', 'när', 'nästa', 'någon', 'någonting', 'något', några', nån', nåt', 'nödvändig', 'nödvändiga', 'nödvändigt', 'nödvändigtvis', 'och', 'också', 'ofta', oftast', 'olika', 'olikt', ‘om', ‘oss', 'på', 'rakt', 'redan', rätt', 'sade', 'sagt', 'samma', 'samt', 'sedan', sen', 'senare', senast', 'sent', 'sex', 'sextio', 'sextionde', 'sexton', 'sextonde', 'sig', 'sin', 'sina', 'sist', 'sista', 'siste', 'sitt', 'sju', 'sjunde', 'sjuttio', 'sjuttionde', 'sjutton', 'sjuttonde', 'själv', 'sjätte', 'ska', 'skall', 'skulle', 'slutligen', 'små', 'smått',

'snart', 'som', 'stor', 'stora', 'stort', 'står', 'större', 'störst', ’säga', 'säger', 'sämre', 'sämst', 'sätt', 'så', 'ta',’ tack','tar', 'tidig', 'tidigare', 'tidigast', 'tidigt', 'till', 'tills', 'tillsammans', 'tio', 'tionde', 'tjugo', ’tjugoen', 'tjugoett',' tjugonde', 'tjugotre', 'tjugotvå', 'tjungo', 'tolfte', 'tolv', 'tre', 'tredje', 'trettio', ‘'trettionde', ’tretton', 'trettonde', ' tro','tror','två',’tvåhundra','under','upp','ur','ursäkt','ut’, utan', 'utanför', 'ute', 'vad', var', 'vara', 'varför', 'varifrån', 'varit', 'varje', 'varken', 'varsågod','vart', 'vem', 'vems', verkligen', 'vet', vi', 'vid', 'vidare', 'viktig', 'viktigare', 'viktigast', 'viktigt', vilka', 'vilken', 'vilket', 'vill', 'visst','väl', 'vänster', 'vänstra', 'värre', 'vår', 'våra', 'vårt', 'än', 'ändå', 'ännu', 'är', 'även', 'åtminstone','åtta','åttio', 'åttionde', 'åttonde',' över', 'övermorgon', 'överst', 'övre', 'nya', 'procent',' ser', 'skriver', 'tog', 'året'

TRITA EECS-EX-2018:441