ERASMUS UNIVERSITY ROTTERDAM
Erasmus School of Economics

Master Thesis
Data Science and Marketing Analytics

Nonlinear Centrality-based Summarization

Name student: Ayan Islamova

Student ID number: 429888

Supervisor: Dr. Eran Raviv

Second assessor: Dr. Erjen JEM van Nierop

Date final version: 24 January, 2021

Abstract

This study aims to assess whether the implementation of the clustering before applying the automatic summarization algorithm would improve the quality of the produced summaries. By splitting the sentences in the news articles into distinct clusters based on the shared features and then applying the LexRank algorithm on each cluster allows accounting for various topics present in the articles. The empirical analysis performed on the 100 CNN news articles revealed that the summaries generated after applying the K-means and Spherical K-means clustering algorithms were closer to the golden standard abstractive summary than the summary that was produced by directly applying the LexRank summarization algorithm. The study concluded that the application of clustering algorithms prior to summary generation leads to the summary of a better quality as compared to the traditional approaches.

Table of Contents

Chapter 1

Introduction

Since the vast amount of information is stored in the form of text data and it is growing daily, the need to process this data to gain valuable knowledge without losing precious time is increasing. This results in a problem of information overload and creates a demand for concise summaries that should contain the key ideas and reflect the main points of a document. Hence, the research community is demonstrating a strong interest in developing new approaches for automatic text summarization. These techniques generate summaries that contain the most important information from the original documents. There are two main methods for the automatic generation of summaries, namely abstractive and extractive (Gambhir & Gupta, 2017). The former method requires extensive natural language preprocessing of the text data and a summary produced with this method will have new sentences that would be constructed in a way that the main idea of the document is preserved. On the other hand, the latter method does not create any new sentences but it simply determines the most informative and valuable sentences from the original document and includes them in the final summary.

While many researchers try to focus on improving the mechanism for automatic summarization, it is crucial to consider how these algorithms capture the importance of certain sentences. There are loads of various factors that contribute to the selection of the most important sentences in the documents, among them are title similarity, sentence position, sentence length, the similarity between all the sentences, etc (Tas & Kiyani, 2007). Some other methods try to use the probability distributions of the sentences and documents over various topics that are present in the documents. Thus, a linear function of thematic features is used for the computations of the scores that eventually determine the likelihood of each sentence belonging to the summary (Hennig, 2009). This method considers different topics of the documents by utilizing the latent semantic analysis, however, the computation of the sentence scores still relies on the linear approach. Focusing on different topics within each document allows us to capture the sentences that may be important to a certain subtopic but not strongly related to the overall topic of the document. Therefore, this paper proposes a new approach to the automatic summarization where incorporation of the nonlinear component would allow to capture the sentences that are considered important for different subtopics of the documents. This can be accomplished by splitting all the sentences in each document into distinct groups

before applying the automatic summarization algorithm. The idea, however, is not to identify an explicit subtopic in each document but rather to group sentences based on the shared characteristics. Hence, sentences in one group might relate to a certain topic. By utilizing the famous clustering algorithm it is possible to partition the sentences in each document in a separate predefined number of clusters or groups where each cluster contains sentences that share similar features. Afterward, the automatic summarization algorithm is applied not on the whole document but on each cluster of sentences which allows selecting a sentence from each cluster to be included in the final summary for the corresponding document. The comparison of these summaries with the global one (without application of clustering) will enable us to arrive at conclusions concerning the quality of the suggested method. Therefore, the research question that will be tackled in this paper has been formulated as follows:

Does the application of clustering algorithms help to improve the quality of summary as compared to the global summary?

Given the significant increase in the number of documents available on the internet over the last years, the demand for quick and efficient summarization techniques is at the peak (Sethi et al., 2017). This is particularly the case for the news articles since their point is to convey the message to the audience as fast and as clearly as possible. If the news article is too lengthy and monotonous, often the reader would not be engaged throughout the article and he/she will stop reading it somewhere in the middle. While automatically generated summaries are very useful for a reader, sometimes these algorithms fail to capture all the important points of the article. At the same time, this is a controversial topic since what one reader may find "important", another reader may consider not worthy of being included in the summary. That is why this paper focuses on finding all the relevant topics within each news article by splitting sentences into distinct groups based on their similarities to capture and convey the information that might have been lost when using traditional approaches. By accounting for the presence of various subtopics in each article in an above-described manner, this thesis constitutes its scientific relevance.

The evaluation of the produced summaries constitutes of calculating distances between the summaries generated post-application of clustering and golden standard summary (human-generated) as well as between global and golden standard summary. The summary for which

the distance to the golden standard summary turned out to be the smallest as compared to the distance between another summary and golden standard one was determined to have a better quality. The analysis and comparison of the summaries of 100 news articles revealed that the summaries generated post-application of clustering algorithms demonstrated a better quality as compared to the global summary.

The following chapter will discuss the most relevant literature related to the topic, followed by the data chapter that will describe the data that has been used in the analysis and preprocessing steps required for the study. The subsequent chapter will focus on the explication of the methods used to perform the empirical analysis, and the ensuing chapter will describe the obtained results. The final chapter will discuss and conclude all the relevant findings of this study, consider the limitations of the research, and provide some suggestions for future research.

Chapter 2

Literature Review

The main challenge that each individual faces when looking for anything online appears to be directly related to information overload which is the situation when a user is not able to process all the available information in the given time interval (Bergamaschi et al., 2010). There is a vast amount of news that becomes available on the internet every day and if they are not accessible in a concise and easily understandable form, then readers would tend to skip the article or do not finish reading the same until the end. Many papers acknowledge the importance of automatic text summarization techniques in tackling the information overload problem. Plaza et al. (2010) mention that automatic summarization can be particularly useful in this setup since it allows readers to obtain an adequate idea of what an article talks about in merely a couple of sentences without the need to complete the whole article. Despite the fact that the research in this area achieved impressive results, there is always a room for improvements.

One of the first endeavors to build an automatic summarization algorithm belongs to Luhn (1958), who defined a technique that would automatically summarize the scientific articles. The assumption was that the significance of a sentence can be derived from analyzing the words

that it contains. The significance of each word, in turn, was determined by looking at the frequency of each word's occurrence in an article. Thus, by calculating the term frequency it is possible to identify the most important, i. e. the most frequent terms which would increase the significance of the sentences that they belong to. Based on this significance factor, sentences with the highest factor score would be selected to be included in the summary.

In the following years, multiple various approaches to automatic text summarization have been proposed in the scientific community. Automatic summarization mechanisms can be classified based on several criteria. For instance, automatic document summarization can be of extractive or abstractive form. In the former approach, the most relevant sentences are selected from the document and they are included in the summary in the same form as they are presented in the original document. Whereas, in the latter approach, the main ideas of the document are rephrased and appear in the summary in a different form. Since the abstractive approach requires advanced natural language processing methods, most of the research in the field is focused on extractive techniques (Gambhir & Gupta, 2017). This thesis accommodates the extractive approach to automatic text summarization.

Extractive automatic summarization techniques can be further categorized based on their underlying algorithm. Thus, there are statistical-based approaches (Ko & Seo, 2008), topic-based approaches (Harabagiu & Lacatusu, 2005), graph-based approaches (Erkan & Radev, 2004), discourse-based approaches (Mann & Thompson, 1988), and approaches based on machine learning (Wong et al., 2008). This thesis focuses on the graph-based techniques, where the documents are represented on a graph. This graph has the nodes that correspond to the sentences or paragraphs that will be extracted from the original document and the edges that represent the similarities between each pair of sentences or paragraphs. The goal of the algorithm is to assign each node a score that would indicate the importance of each node, after which all the nodes are ranked and the most important nodes which at the same time are sentences or paragraphs will be included in the summary.

In one of the early papers where the graph-based approach has been utilized for automatic document summarization, the authors mention the tremendous benefit of the proposed technique in a battle with information overload (Salton et al., 1997). Their algorithm generates summaries by extracting the most relevant paragraphs from the original documents. Since the

method relies on the degree centrality, once the graph with the nodes corresponding to the paragraphs and the edges is constructed, a certain similarity threshold is defined. Based on this threshold, all the connections between the paragraphs that fall under the threshold are excluded. After that, the algorithm counts the number of remaining edges that are connected to each of the nodes on a graph. A paragraph that has the most paragraphs connected to it, indicating that these paragraphs are similar, will be considered as more important and hence, it will be included in the final summary. This summarization approach has been applied to fifty articles from the encyclopedia and it has been evaluated by comparing the produced summaries with the human-generated ones. The proposed technique demonstrated a substantially better performance as compared to the approach of selecting random paragraphs from the articles.

After a few years, inspired by Google's PageRank algorithm (Brin & Page, 1998) that is used to rank the webpages, two new ranking algorithms for automatic text summarization were introduced in the academic world. The algorithms TextRank and LexRank were proposed by Mihalcea and Tarau (2004) and Erkan and Radev (2004) respectively. Both ranking algorithms incorporate an important alteration to the algorithm proposed by Salton et al. (1997). By taking a step further, both algorithms determine the significance of each node not only based on the number of nodes that it connects to as proposed by Salton et al. (1997), but also by the significance of the nodes that it is connected to. This provides a considerable improvement, however, since the algorithm performs a discretization operation while calculating the LexRank scores, this results in the loss of some information. To account for this, another improvement was suggested by Erkan and Radev (2004) where while computing the LexRank scores an algorithm also multiplies the obtained values of the linking sentences by the links' weights. This alteration results in the adjusted version of LexRank for weighted graphs and it is named continuous LexRank. Since this paper does not attempt at improving the underlying automatic summarization algorithm, but rather applying the existing algorithm on the different clusters of sentences, a continuous LexRank algorithm was selected to produce the final summaries in this thesis.

Since this paper attempts at applying the automatic summarization algorithm to the clusters of sentences for each article, it is paramount to consider the way partitioning is performed. The topic of document clustering is a well-studied matter and to this day there are multiple techniques available. One of the popular clustering algorithms also known as K-means

clustering attempts at partitioning the data points such that similar patterns are observed in the same clusters and the different patterns are present when looking at two different clusters (Alsabti et al., 1997). This algorithm has proven to generate good results by partitioning the data in clusters effectively. However, the pitfall of the method is the fact that implementation of the technique can be very computationally expensive and time-consuming especially when working with large data sets. Another inconvenient aspect of the k-means algorithm is that the number of produced clusters has to be specified in advance and determining the optimal number of clusters is a laborious task. Selecting a large number of clusters might complicate the analysis and interpretation of the obtained results, whereas, selecting a very small number of clusters may result in a loss of valuable information (Xu & Wunsch, 2005).

Despite the fact that the K-means clustering algorithm has been proven to be an efficient technique since it utilizes the Euclidean similarity measure it is often not considered to be a suitable choice when it comes to document clustering (Dhillon et al., 2002). Another similarity measure that has proven to be effective for document clustering named Cosine similarity, employs the cosine of an angle between two vectors of documents. By integrating the Cosine similarity measure into the standard K-means algorithm, we arrive at a new Spherical K-means algorithm. The name of this algorithm comes from the fact that the calculations are performed on the document vectors that are projected on a unit sphere. This clustering algorithm has proven to be one of the most efficient ones when it comes to document clustering (Zhong, 2005). Therefore, this thesis will utilize both the standard K-means and Spherical K-means algorithms for the partitioning of the sentences in each article into different clusters based on the shared features.

Due to the unavailability of a human reader who ideally could compare produced extractive summaries against the abstractive human-generated summary this paper had to accommodate an alternative approach for evaluation. Therefore, to compare the summaries in this study all summaries are represented as vectors and the distances between them are computed. One of the most famous techniques for obtaining a vector representation of the documents is called Doc2Vec and it was introduced by Le & Mikolov (2014). The proposed algorithm is nothing but an extended version of the earlier introduced Word2Vec model (Mikolov et al., 2013) that is used for mapping the words to embedding vectors. This model produces the continuous word vectors that are also known as word embeddings, where each word embedding captures the

contextual meaning of a corresponding word. The method relies on the assumption that the meaning of a word is defined by its context or the words surrounding it. Under the Word2Vec approach, the authors propose two models both of which rely on artificial neural networks. One of the models named Continuous Bag-of-Words (CBOW) attempts at predicting a focal word based on its context (surrounding words) by training a neural network. In another model named Skip-gram, the goal is to predict the surrounding words from the focal word. Both models are trained by the artificial neural network with a hidden layer. This embedding representation of words can capture the complex relationships between words and it is proven to have a better performance than other methods for obtaining the continuous word vectors as concluded by the authors.

In the context of automatic text summarization, however, the goal is to arrive at the embedding representation of not only words but the whole document, i. e. produced summary. Thus, the focus lies in the Doc2Vec method where the context of a word is driven not only by the words surrounding it but also by a sentence or a paragraph that this word belongs to. Therefore, the paragraph vectors also assist in the prediction of the next focal word in a paragraph given other words in the context (Le & Mikolov, 2014). Following this intuition, the Doc2Vec architecture is similar to the Word2Vec one with the modification where the IDs of the sentences or paragraphs that contain the focal word are treated as an extra token in the model. Similarly to the Word2Vec approach, the Doc2Vec method also consists of two models namely the Distributed Memory Model of Paragraph Vectors (PV-DM) and the Distributed Bag-of-Words Model (DBOW). In this study, the preference goes to the former model since it also accounts for the word order in the sentences, unlike the DBOW model that relies on the bag-of-words assumption.

Chapter 3

Data

The data set used in this research is a set of CNN news articles and their highlights[1]. The initial dataset consists of thousands of files each of which contains a news article and highlights which are a few sentences of abstractive summary for each article. For the analysis conducted in this paper, a subset of 100 articles together with their highlights were randomly selected and stored in a separate excel file manually. Generally, the more observations are present in the analysis

[1] The dataset was obtained from https://github.com/deepmind/rc-data

the more computationally expensive it becomes. Considering this in addition to the fact that this research accommodates multiple methods it has been decided to use a subset of the data. The chosen number of observations offers a good balance between the computational efficiency and the opportunity to arrive at the meaningful conclusions when evaluating the quality of the produced summaries.

The dataset consists of 100 observations and 3 features: a unique article ID, article itself, and highlights summary. To get some basic information about the data, articles, and highlights were tokenized into sentences and afterward, the mean, median, and mode were calculated for each of the two columns. If a mean value shows the average number of sentences in the article, a median would be a value that is located at the midpoint of the whole frequency distribution of all sentences and a mode would be the value that appears in the distribution the most. Table 1 indicates that on average each original article contains around 38 sentences whereas the median and the mode for all articles are 35.5 and 32 respectively. For the highlights, however, the average number of the sentences in each highlight turned out to be around 4 sentences with both median and mode values being 4. These insights are useful in making some important decisions in the future sections of the paper, for instance, in determining the optimal number of sentences in the produced summaries.

Table 1: The number of sentences in Articles and in Highlights

	Mean	Median	Mode
Articles	37.79	35.5	32
Highlights	3.71	4	4

The following section will address the data preprocessing in R, and later on the data preprocessing steps in Python will be discussed.

To answer the research question, that is "*Does the application of clustering algorithms help to improve the quality of summary as compared to the global summary?"* and compare the qualities of different summaries, first, those summaries need to be produced. Since the summarization algorithm which is used in this research is applied to each article on a sentence level each article is stored in a separate data frame where each row corresponds to a single sentence. Before calculating the Lexrank scores for each sentence the arguments specified in the function perform some basic text cleaning. Such cleaning removes all the punctuation and

numbers from the text and sets all the words to the lower case. Furthermore, it removes all the commonly used stopwords in the English language from the SMART information retrieval system[2]. Lastly, it performs stemming[3] which is a process of cutting down the words to their base or root form. As an example, the words "argue", "argued", "argues", and "arguing" would become "argu" which is the stem of all these words. The stemming is performed with the "wordStem" function from the *SnowballC* library and it uses Porter's stemming algorithm which is one of the most popular algorithms for word stemming (Porter, 2001).

Another technique that has been used in this paper called clustering requires additional data preprocessing steps. Since clustering is applied not on the whole data but each article separately the first step similarly to the one described above includes splitting each article into the sentences and storing them in a separate data frame with the corresponding sentence ID. The algorithm takes as an input a Document Term Matrix (DTM) which is a matrix that shows how frequently each term occurs in a collection of sentences. In this matrix, each row corresponds to a sentence from an article and each column corresponds to a term. To generate a DTM first a data frame with all the sentences from one article is transformed into a corpus of sentences and afterward, this corpus is converted into a DTM. Each value in the given matrix represents a Term Frequency (TF) which is weighted by the Inverse Document Frequency (IDF). TF corresponds to the number of times a certain term occurs in a sentence. Nevertheless, a certain term could be used very frequently and it can occur in many sentences in the article. In this case, the term would not be very informative and valuable. To solve this issue the frequency of a term should be considered in relation to the length of the sentence. This is done by multiplying the TF score by the term's IDF score and it can be calculated as

$$IDF(w_i) = log\ \frac{N}{n_i},$$

where N is the total number of sentences and term w_i occurs in n_i of them (Robertson, 2004). In this research, N would correspond to the total number of sentences in each article and it would vary with every article in the data set. This weighting method ensures that the terms which occur more frequently are not weighted too heavily. Furthermore, the very common terms, for instance, prepositions, articles, and conjunctions will get a lower weighting. In this process, some of the simple text transformations are applied to the sentences. All the numbers, punctuation, and common English stop words are removed. In addition, all the words are

[2] https://jmlr.csail.mit.edu/papers/volume5/lewis04a/
[3] https://dergipark.org.tr/en/download/article-file/392456

transformed into the lower case, and stemming is applied to all the words in each sentence. After the DTM is constructed all the rows which contained only zeros were removed because they contain empty sentences. This was done to avoid problems with the clustering algorithm that cannot be applied to the empty documents which in this case are represented by sentences. The last step comprises the normalisation of term vectors by applying a "scale" function on the DTM. The values in the DTM correspond to the continuous numeric features and the clustering algorithm relies on the average of these numbers. Hence, it is important to normalise these values before applying the clustering algorithm.

Moving on to the data preprocessing in Python, once the data is loaded, it is stored in the Pandas data frame from the Pandas library. The text data preprocessing and cleaning steps here consist of conversion of the text into the ASCII (American Standard Code for Information Interchange), removing punctuation, numbers, and extra white space between the words. Next, each article is tokenized on the word level and this becomes an input for the stemming function which leaves only a stem of each word using Porter's stemming algorithm. After the stemming is performed all the words are concatenated back into one string for each article.

Chapter 4

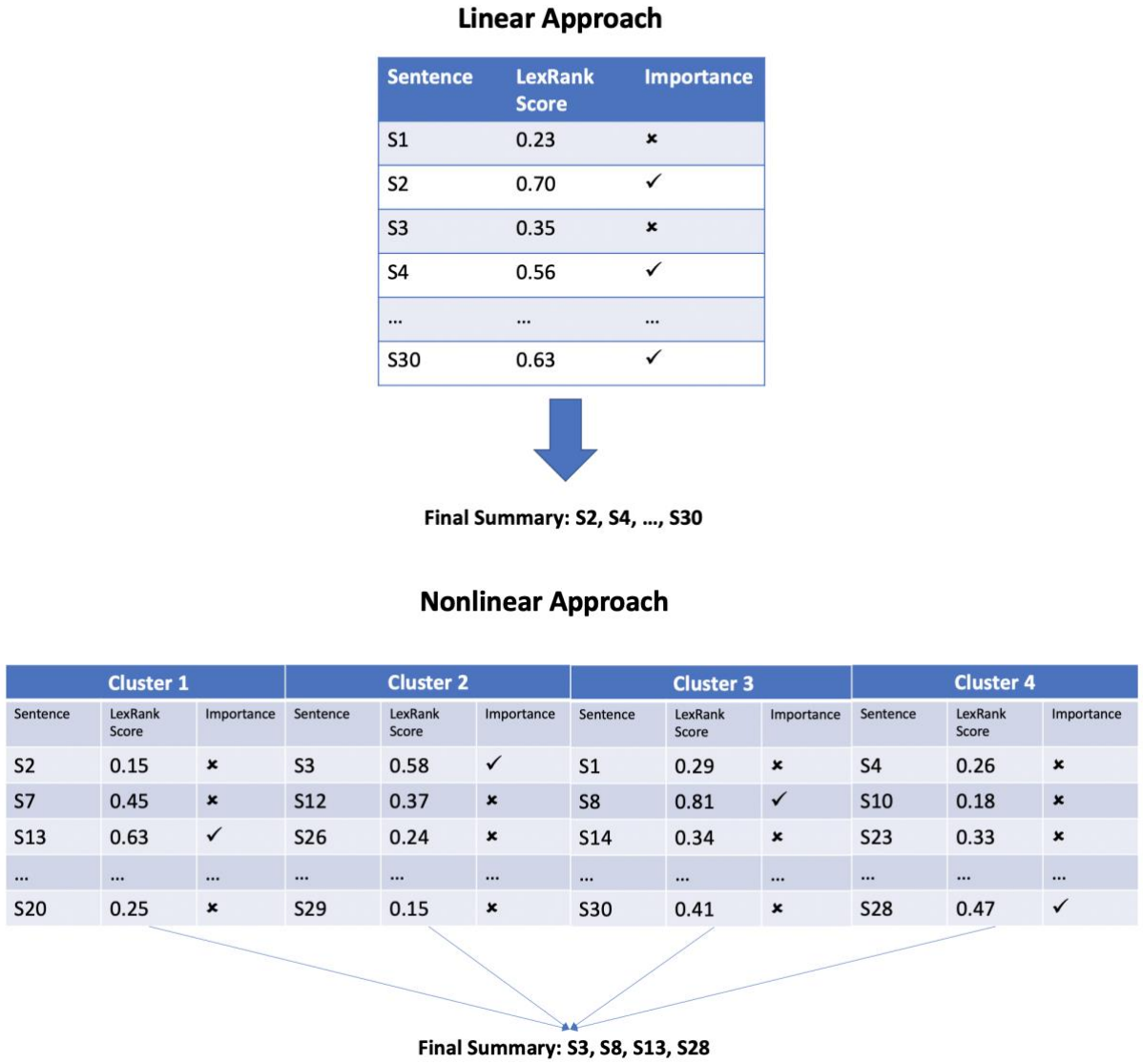
Methodology

There are several techniques that have been applied in this research. The purpose of this thesis is to compare various summarization methods that were applied on the same news articles and to determine the ones that perform better than others. Overall, four different summarization methods will be compared and evaluated. One summary was already provided together with the data and it is the only summary which is of abstractive form. All of the remaining three summaries are extractive in their nature and they have been generated using the LexRank algorithm. The difference between these three summaries is that one summary was generated by applying the Lexrank algorithm on the whole article and the other two summaries were produced by first splitting the sentences in each article into different groups or clusters and then applying the LexRank algorithm. Hence, in the case of the summary that is generated by applying the LexRank on the whole article, each article is treated as one single cluster by the algorithm. This summary is called "Global Summary". For the other two extractive summaries, before applying the LexRank first the clustering technique has been used to split the sentences

in each article into different clusters. The key added value here is that if the summarization algorithm is applied to the whole article, each sentence in that article would get a LexRank score based on the sentence centrality, i. e. centrality of a sentence represents how closely that sentence is related to all other sentences in a certain cluster (in this case the whole article is one cluster). However, when the sentences in an article are split into different clusters before applying the LexRank algorithm, it allows capturing the importance of sentences that are not central to the whole article but which are central to a specific cluster. Thus, the sentence centrality is no longer the only factor that dictates the importance of a sentence for an article that it belongs to. This way, sentences that would have gotten a low LexRank score and would have appeared at the bottom of the list get higher chance of being selected for the final summary. This nonlinear element allows us to account for the sentences that are important to different clusters and it allows us to add some of these sentences describing different topics in the final summary.

The differences between the linear and nonlinear approaches are depicted in Figure 1. In this example, the algorithm is applied on the article that consists of 30 sentences. In the linear approach, a summarization algorithm is applied on the whole article, and each sentence in that article is assigned to the corresponding LexRank score. The sentences with the highest LexRank scores get selected to be included in the final summary for this article. On the other hand, in the nonlinear approach, all 30 sentences first are split into four clusters, based on their similarities, after which the summarization algorithm is applied on each cluster separately. Each sentence in each cluster is assigned to the corresponding LexRank score. After that, the sentence with the highest LexRank score is selected from each cluster and these four sentences are included in the final summary for the given article. This example clearly illustrates the general idea of the proposed method. Here, for instance, sentence number 3 is only considered important for cluster 2, thus, it would not appear in the final summary in the case of the linear approach.

Figure 1: Linear versus Nonlinear Approach to Automatic Text Summarization



To split the sentences in each article into the groups different clustering techniques have been accommodated in this research. The first one is named K-means clustering and the other one is named Spherical K-means clustering. To understand how these algorithms select sentences and place them in the same clusters a more detailed description of these methods will be provided in the coming sections of the paper.

K-means Clustering

Since the idea of splitting sentences into different groups relies on the fact that each group should contain sentences that have something in common an appropriate method has to be selected. One of the popular approaches for segmentation of the data is a so-called K-means

clustering. This method attempts to partition the data points (in this case sentences in each article) into a certain number of groups or clusters in such a way that the sum of squares from these data points to the centers of a cluster (centroids) is minimized. The goal of the algorithm is to find a local optimum which is the case when no further movement of a data point from one cluster to another would result in the lower value of the within-cluster sum of squares (Hartigan & Wong, 1979).

The algorithm of K-means clustering works as follows. The number of clusters has to be specified in the function before the partition occurs. Once this is done, a random point gets selected as a centroid for each and every cluster. Then each sentence gets assigned to the nearest centroid. After all sentences in an article are assigned to their centroid, the algorithm calculates the sum of all distances to the centroid. This is when the centroids change their original randomly chosen location and move to the average distance sum of all assigned sentences. Naturally, when the centroids of the clusters move, some sentences will switch from one cluster to another. The algorithm will keep adjusting the boundaries of clusters as well as the mean centroids until the within-cluster sum of squares is at the minimum level. Once no further improvement can be achieved the algorithm stops computations and all the sentences stay in their cluster.

Due to the fact that the computation of the mean deviation is required, this approach relies on the Euclidean distance measure (Chapman & Feit, 2015). The Euclidean distance corresponds to the root sum-of-squares of differences between sentences. The Euclidean distance between two sentences x and y represented as vectors in each dimension from i to n is calculated as

$$d(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}.$$

An important and difficult choice that has to be made when running this algorithm is the number of clusters since it has to be specified in advance. This choice usually depends on the domain of a study and several values have to be tried to find the optimal solution. It is crucial to consider the computational cost of the implementation of K-means clustering per iteration since the algorithm needs to calculate the centroids and distances (Alsabti et al., 1997). Furthermore, this process is repeated for all the sentences in each of a hundred articles in the

dataset. Hence, the implementation of the K-means technique can be very costly in terms of computation time.

Coming back to the problem of specifying the number of clusters, the decision regarding the optimal number of clusters (k) has to be taken. Based on one of the solutions of trial and error, several numbers of clusters have been tried out, however, this procedure turned out to be computationally too intensive in terms of time. Thus, it has been decided to focus only on one number of clusters for both clustering techniques that have been applied in this research. After separating all the sentences in an article into different clusters, the sentences that got the highest LexRank scores were selected to be included in the final summary for that article. Due to the fact that the original summary which was provided together with the articles on average has around four sentences, along with the median and mode values of four sentences it has been decided to split the sentences in each article into four clusters. This means that the final summary for each article will contain four sentences. After the procedure is completed, the sentences that received the highest LexRank scores in each cluster were selected to be a part of the final summary for the corresponding article. In some cases, it could happen that after the partitioning of the sentences into clusters the clusters end up containing only one sentence. If this is the case, then it has been decided to directly include that one sentence into the final summary.

Spherical K-means Clustering

Another technique that was utilized to split the sentences in each article into four clusters is called Spherical K-means clustering. It is useful to try different approaches to compare and determine how various ways of sentences partitioning affect the quality of the final summary. One benefit of this algorithm is its great ability to handle sparse matrices (Kwartler, 2017). This is particularly useful for our research since the DTMs that are used as input to the algorithms are very sparse.

Furthermore, it has been established that standard K-means clustering which utilizes Euclidean similarity measure might overrepresent long documents that contain large aggregate term weights (Buchta et al., 2012). Thus, to alleviate and account for the effect of different document lengths Dhillon and Modha (2001) proposed to accommodate a Cosine similarity instead of Euclidean similarity, which is equivalent of using Euclidean similarity after projecting the

vectors of features onto the unit sphere. Cosine similarity measure corresponds to the cosine of the angle between the sentences vectors after those vectors have been normalized and it can be calculated using the formula below:

$$cos\ (\vec{v}, \vec{w}) = \frac{\vec{v} \cdot \vec{s}}{|\vec{v}||\vec{s}|} = \frac{\sum_{i=1}^{N} v_i \times s_i}{\sqrt{\sum_{i=1}^{N} v_i^2} \sqrt{\sum_{i=1}^{N} s_i^2}},$$

where v and s are two sentences vectors in N dimensions. The values of the Cosine similarity range from -1 to 1. The Cosine similarity of -1 between the two sentences would indicate that the sentences have the opposite meaning, whereas the Cosine similarity value of 1 would indicate that the two sentences are exactly the same. If the value of Cosine similarity between two sentences corresponds to 0 it would suggest that these two sentences are not related.

Consequently, the problem that Spherical K-means is trying to solve is the minimization of

$$\sum_{i} (1 - cos\left(x_i, p_{c(i)}\right))$$

for all assignments c of sentences i to the cluster IDs $c(i) \in \{1, \ldots, k\}$ and for all centroids (prototypes) $p_1, \ldots, p_k$ in the same feature space as the feature vectors x_i that represent the sentences (Buchta et al., 2012).

Spherical K-means can be implemented with the *skmeans* function from the *skmeans* package. This function takes a normalized DTM as an input and a few other parameters need to be specified for the partition to occur. Similarly to the standard K-means problem, the desired number of clusters needs to be prespecified. For the reasons described in the above sections of the paper in this case it was chosen to split the sentences in each article into four clusters. Having the same number of clusters in both clustering techniques would allow performing a fair comparison between the final summaries which are produced with two different methods. There is a new parameter here, however, which is not present in the case of the standard K-means approach. This is a parameter *m* which controls the fuzziness of the obtained clusters. Controlling this parameter allows for the border of a cluster sphere which is not defined in a concrete way. The minimum value of *m* is always 1 and this indicates a hard partition that is not allowing for any fuzziness between the borders of clusters. The fuzziness of the cluster

borders will keep increasing with the value of *m*. The last two control parameters are *nruns* and *verbose*. The first one makes sure that the function will rerun the building of a model a predefined number of times. This is similar to performing a cross-validation in some other machine learning techniques which helps to ensure that the results of the model are stable enough. The second control parameter, once set to true, would print the progress of the computations to the console while the model is building to provide an indication of whether the model has frozen the computer. This is particularly useful when building very computationally expensive models.

The Spherical K-means function accommodates many various computation algorithms, however, the only algorithm which can be used when doing a soft partition is called *pclust*. This algorithm allows to perform fixed-point runs several times and this number can be controlled with the parameter *nruns* that was described in the paragraph above. As in the case of this paper, if the starting values are not explicitly defined the algorithm will build a model as many times as specified in the parameter *nruns* and it will use the default initialization. The partition that is returned by the model would be the first one that corresponds to the smallest value of a criterion in all the performed runs (Buchta et al., 2012).

LexRank Algorithm

The authors of an algorithm that was used to generate the summaries of the articles named LexRank got their inspiration from another algorithm - PageRank which is used to determine the importance of the webpages. The difference is that the LexRank algorithm is commonly used to rank the chunks of text in the documents. While many factors could indicate the importance of a sentence in the article such as the position of a sentence in an article or an overall frequency of the words in a sentence, the goal of the LexRank algorithm is to assess the centrality of each sentence from the lexical point of view. Unlike in some other graph-based approaches to text summarisation where an algorithm is based on the degree centrality and importance of a sentence is established by the number of sentences that it connects to on a graph, the algorithm of LexRank determines the importance of a sentence differently. A substantial difference in selecting the most valuable sentences with the LexRank algorithm is that it does not only look at the number of sentences that are connected to a certain sentence but it also considers the importance of those sentences.

An underlying assumption is that the sentences in an article that would be defined as more central or more salient are the ones that are similar to the majority of other sentences in an article. Thus, it is crucial to define how the similarity between the sentences is measured. To understand how similarity between two sentences can be measured first each sentence is represented as a vector with N dimensions where N corresponds to the number of all the words that can be encountered in a target language. This model is called bag-of-words since each word in a sentence is represented by a separate dimension and it has no correlation with the other words in a sentence. In this vector representation of a sentence, the value of the dimension for each word that is occurring in a sentence would correspond to the number of its occurrences in a sentence multiplied by its IDF. Given this, the similarity between two sentences in an article is measured by the cosine between the vectors which represent these two sentences:

$$idf - modified - cosine(x, y) = \frac{\sum_{w \in x,y} tf_{w,x} tf_{w,y} (idf_w)^2}{\sqrt{\sum_{x_i \in x} (tf_{x_i,x} idf_{x_i})^2} \times \sqrt{\sum_{y_i \in y} (tf_{y_i,y} idf_{y_i})^2}}.$$

In this formula $tf_{w,s}$ measures how many times the word w occurs in the sentence s.

A collection of articles can be represented by a so-called cosine similarity matrix where each value matches the similarity between all the corresponding sentence pairs. Another way to represent this matrix is to put it in a weighted graph form where each node corresponds to the cosine similarity between two sentences.

The problem that arises when the centrality of a sentence in an article is determined by only looking at the number of sentences that it connects to (in a graph representation, the number of connections between the edges which represent all the sentences in an article) is that it ignores the fact that some relationships may not carry as much importance as others. As a result, several unimportant sentences may vote for each other and get selected to be included in the final summary which is not the desired outcome. This is where the LexRank algorithm makes an adjustment by considering the nature of those votes and by taking the centrality of the voting nodes into account by weighting every node. Hence, it is considered that every node has its own centrality which is distributed to the neighboring nodes. This can be noted in the following manner:

$$p(u) = \sum_{v \in adj[u]} \frac{p(v)}{deg\ (v)},$$

where $p(u)$ shows the centrality of node u, $adj[u]$ represents a set of nodes which are adjacent to node u, and $deg\ (v)$ corresponds to the degree of the node v.

This formula is a base for an underlying algorithm of the PageRank which is widely used in determining the prestige of the webpages and it can be depicted in the following way:

$$p(u) = \frac{d}{N} + (1 - d) \sum_{v \in adj[u]} \frac{p(v)}{deg\ (v)},$$

where N corresponds to the total number of nodes in the graph, and d is a so-called "damping factor" which regulates the convergence of the method. The damping factor incorporates the probability of jumping from one node to another random node in the graph into the model (Mihalcea & Tarau, 2004).

The algorithm that calculates the degree centrality and the LexRank scores by using a prespecified threshold does a binary discretization on the cosine matrix. This means that the algorithm will look at the obtained value in a cosine matrix and if the value is above the threshold then it gets a score of 1. Contrary, if the algorithm determines that the value in the cosine matrix is below the threshold then it gets a score of 0. This process as other discretization operations results in information loss. Thus, another improvement that can be achieved over the LexRank is to incorporate the strength of the similarity links (Erkan & Radev, 2004). This addition will modify the similarity graph as well by using the values in the cosine matrix directly, hence, obtaining a denser and at the same time weighted graph. This addition results in the altered version of LexRank for the weighted graphs and it is called a continuous LexRank which can be represented as:

$$p(u) = \frac{d}{N} + (1 - d) \sum_{v \in adj[u]} \frac{idf\text{-}modified\text{-}cosine(u,v)}{\sum_{z \in adj[v]} idf\text{-}modified\text{-}cosine(z,v)} p(v).$$

With this adjustment, when the algorithm computes the LexRank scores it also multiplies these values of the linking sentences by the weights of those links. The weights of the links get normalized by the sums of rows.

Evaluation and Comparison of Summaries

After all the summaries are generated an obvious next step in the analysis is to assess their quality and to identify which summary contains the most relevant information for a reader. In an ideal scenario, the quality of summaries should be assessed by the human reader who can distinguish the most important sentences in each article and decide whether they should have been added to the summary. Moreover, a human reader would be able to compare extractive summaries produced in this study and abstractive summary provided with the data. However, due to the unavailability of a human reader as well as the fact that it would simply take a vast amount of time to read and assess the summaries of hundred articles another method of assessment and comparison of extractive and abstractive summaries is required. One solution that is proposed in this paper is to look at the distance between different summaries of each article and determine which of the generated summaries is located the closest to the golden standard summary Highlights. It is important to keep in mind that the golden standard summary is of an abstractive form, unlike the three remaining extractive summaries which were produced by the LexRank algorithm.

To measure the distances between the summaries first of all these summaries have to be represented as vectors. An underlying idea is that each word can be mapped to a vector of real numbers and having this word vector representation allows computing the distances between words in an embedding space. However, to answer the research question in this paper, a comparison between the whole summary for a certain article with a different summary for the same article is required. This means that not every single word must be represented as a vector of numbers but the whole summary which contains four sentences needs to be converted into an embedding. An algorithm that transforms the text in the documents into vectors is called Doc2Vec. A detailed description of this algorithm will be presented in the following section.

Doc2Vec Algorithm

One of the most popular tools in natural language processing for obtaining the vector representation of documents is a model called Doc2Vec that was introduced by Le & Mikolov (2014). This algorithm was proposed as an extension of Word2Vec (Mikolov et al., 2013) which works similarly but its purpose is to map the words into an embedding space. Both algorithms rely on the Distributional Hypothesis (Harris, 1954), which is the core of many NLP techniques. This hypothesis relies on the assumption that the meaning of a word is driven by its context and two words would be considered similar if they share similar contexts. There are two models in Doc2Vec that can transform the documents into embeddings namely the Distributed Memory Model of Paragraph Vectors (PV-DM) and the Distributed Bag-of-Words Model (DBOW). Since the latter model relies on the bag-of-words assumption it does not take the order of words into account. For this research, however, it is crucial to consider the word order in the sentences since the entire summary with four sentences will be mapped into the vector space. Therefore, in this study, the preference goes to the PV-DM model. Since the PV-DM model was introduced as an extension of the Continuous Bag-of-Words (CBOW) model of Word2Vec, the paper will first discuss the architecture of this model and then it will focus on the PV-DM model.

In essence, the goal of the CBOW model is to train a feedforward neural network so that it can predict a focal word given the other words in a context. A feedforward neural network model tries to imitate the process which is similar to the way a human brain operates by introducing a network of interconnected neurons that help to approximate the complex non-linear relationships in the data. These neurons are captured in several layers namely an input layer which represents every neuron as a predictive feature of the model, the hidden layers where output is produced by calculating the weighted inputs, and an output layer that produces the response value given the input values. Hence, each layer of the network contains the interconnected neurons or nodes that are weighted with the neurons in the prior and subsequent layers. The neurons in the input layer contain the raw input values, whereas, the neurons in the hidden and output layers take as an input the weighted sum of the previous layer's inputs and pass the obtained results through a predefined activation function to generate the output. This output can become an input of the next hidden layer or in case there is no other hidden layer remaining it can be considered as the final output of the model (Goodfellow et al., 2016).

Mathematically, for any hidden or output layer with the prior layer containing n neurons and the current layer containing v neurons the problem can be generalized to:

$$\left(\sum_{i=1}^{n} k_{i,v} * x_i\right) + k_0 = y_j,$$

where $k_{i,v}$ corresponds to the weight connection between the neurons in the previous and current layers, x_i is either the raw input in the input layer or the value obtained from the previous layer, k_0 is added to account for the bias and to adjust the output value with the weighted sum of the inputs to the neurons, and y_j represents the activation value for the neuron j in the current layer. Once all y_j values are calculated an activation function is used to compute the activation values in the current layer. The activation function is associated with each neuron in the network except for the neurons in the input layer as it only contains the raw input features. Hence, after calculating every y_j value and applying the activation function for each subsequent layer eventually the response values at the output layer are obtained. At the next step, a predefined loss function is used to calculate the error which compares the obtained response value with the actual observed response value. Afterward, this prediction error is propagated back through the network using the same weights that were used to forward propagate the input values. This algorithm is called backpropagation and it helps to identify every neuron's contribution to the overall loss of the network. When the error is propagated back a gradient is calculated by taking the partial derivatives with respect to every single weight in the network. The result is a vector that points in the direction in which the weights in the matrices will be updated to arrive at the minimum value of the loss function. Thus, all the weights in the network get updated after each training instance is forward and backpropagated and this process is repeated for the fixed number of epochs, i. e. training cycles until all the training instances go through the network. This iterative method that enables the minimization of the loss function is called stochastic gradient descent and it allows us to find different patterns in the network in each epoch to arrive at a better generalization performance. To avoid the model convergence at the suboptimal values of the weights every time the weights get updated the update gets scaled down by multiplying it by the learning rate for which the values range between 0 and 1. Nevertheless, one has to be careful when selecting the value for the learning rate as very low values may lead to a longer training time. The common way is to select the higher value for the learning rate at the beginning of the training and keep decreasing it with every epoch to avoid overshooting the minimum.

In the CBOW architecture, the feedforward neural network is trained to predict the focal word from the other words in its context. However, the embeddings of the words are stored in the weight matrix located between the input and the hidden layers. This matrix captures the relationships between the input words as well as their contextual meaning. Before the neural network gets trained on all the instances a context of the focal word has to be defined. The context for each word is defined by the window size which corresponds to the number of preceding and subsequent words around the focal word. This logic can be represented as illustrated in Figure 2. Figure 2(a) visualizes the general CBOW architecture where it can be seen that the goal is to predict the target word $w(t)$ by using the context words $w(t-2), w(t-1), w(t+1), w(t+2)$ as the predictors. In this illustration, the window size is set to 2 since the context is defined by the two preceding and two subsequent words around $w(t)$. Figure 2(b) depicts the way training instances get generated where each word is selected as a focal word at a certain point in time and it forms the output value whereas the defined context words are used as the input values for the neural network.

Figure 2: The Continuous Bag-of-Words Model (CBOW)[4]

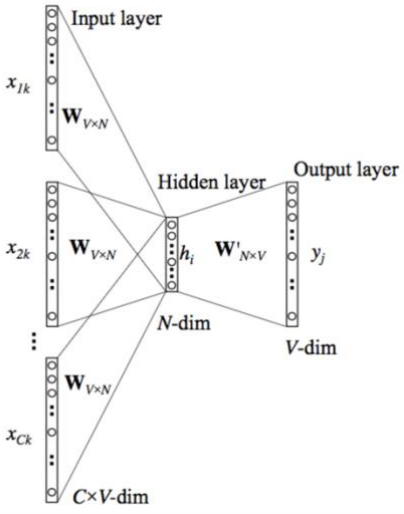
(a) General Structure of CBOW Model *(b) Example of construction of the training instances*

The architecture of the neural network that utilizes the CBOW model is displayed in Figure 3. This architecture represents a slightly modified version of the feedforward neural network that was described in the paragraph above. The figure illustrates that the inputs of the model are all the context words that are represented as x_{1k}, x_{2k}, x_{Ck}. In this architecture, the input layer of the network comprises the C one-hot encoded row vectors with the dimension $1 \times V$ with C representing the total number of the context words and V representing the size of a

[4] Source - https://towardsdatascience.com/nlp-101-word2vec-skip-gram-and-cbow-93512ee24314

vocabulary for the whole corpus. One-hot encoding allows us to label the context words with 1 and the remaining words in the vocabulary with 0 in every vector. Each vector is multiplied by the matrix with the weights represented by $W_{V\times N}$ with the dimension $V \times N$ where N stands for the size of the embedding vector. This weight matrix maps the input X to the hidden layer and this procedure allows us to obtain C vectors each of which has a dimension $1 \times N$ by extracting the rows from the $W_{V\times N}$ matrix that correspond to the positions of the context words in the entire vocabulary. Thus, each word is represented by a vector of continuous values or latent dimensions and the goal of the model is to learn this representation. Moving forward, the hidden layer of the network comprises a $1 \times N$ vector that is obtained by taking the average of all C embedding vectors. Since the activation function of the hidden layer is an identity function, the neurons in the hidden layer copy the weighted sum of inputs to the next layer in the network. In the next step, the weight matrix $W'_{N\times V}$ maps the values obtained in the hidden layer to the final output later with the dimensions $V \times N$. Only at this stage a non-linearity is introduced by applying the softmax activation function in the output layer and by producing the softmax values as an output. Softmax function converts the values in the vector by producing a probability distribution where each neuron in the output would represent the probability of a certain word corresponding to the context word that was used as an input of the model.

Figure 3: The Architecture of the Continuous Bag-of-Words Model (CBOW)[5]



Mathematically, provided the sequence of training words $w_1, w_2, w_3, \ldots, w_T$, the goal of the model is to maximize the average log probability

[5] Source - https://towardsdatascience.com/introduction-to-word-embedding-and-word2vec-652d0c2060fa

$$\frac{1}{T}\sum_{t=k}^{T-k} log\ p(w_t|w_{t-k}, \dots, w_{t+k}),$$

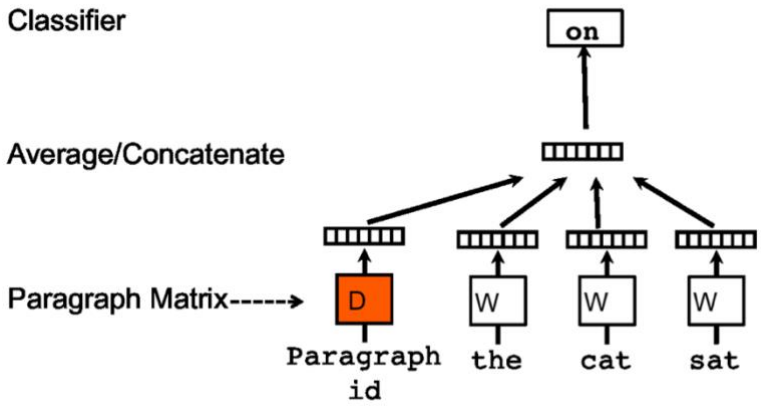
where T corresponds to the total number of terms in the corpus and k is the window size. The probabilities are obtained by applying the multiclass classifier that is a softmax function which is defined as

$$p(w_t|w_{t-k}, \dots, w_{t+k}) = \frac{e^{y_{w_t}}}{\sum_i e^{y_i}}.$$

In this equation, $e^{y_{w_t}}$ is the exponent of y_{w_t} which is the pre-activation value of the neuron in the output layer for the focal word w_t and y_i are the pre-activation values of the neurons for every remaining word in the vocabulary. Thus, in this architecture, a neural network is trained to maximize the probability of predicting the focal word given its context. However, the main goal for this study remains to obtain the $W_{V\times N}$ matrix as its rows correspond to the embeddings vectors of all the terms in the vocabulary.

Now that the architecture of the CBOW model is clearly outlined it is time to discuss the architecture of the PV-DM model since it was introduced as an extension of the CBOW model. The main difference of the PV-DM model is that it does not only use the context words to predict the focal word but it also assumes that the paragraph itself can be used as a predictor of the focal word. This can be implemented by treating the paragraph itself as a part of the context and thus by mapping it to the unique embedding vector in a similar way it is conducted for the words. To apply this into practice, a unique paragraph ID is introduced for every paragraph in the corpus. These unique IDs will be treated as additional terms in the vocabulary. The framework for learning the paragraph vector is illustrated in Figure 4. It can be seen from the figure that the only change from the CBOW model is the additional paragraph ID that can be mapped to an embedding vector via matrix D. The embedding vector of a paragraph represents the information that is missing from the current context and it can act as a memory of the paragraph's topic. The context of a focal word in the PV-DM model is defined as C preceding words plus the paragraph ID of a paragraph that this word belongs to. This helps to capture the word order in the model. Therefore, for each focal word in the corpus, there will be $C + 1$ training instances initialized. Consequently, in this framework, a neural network model will be trained to predict the next word in a paragraph given the previous words plus the paragraph ID.

Figure 4: General Structure of the Distributed Memory Model of Paragraph Vectors (PV-DM)[6]



This alteration results in a slight change in the neural network architecture that was introduced for the CBOW model. For the PV-DM model an input layer will consist of the C one-hot encoded vectors with the dimension of $1 \times V$ for the terms that come before the focal word, and one one-hot encoded vector with the dimension of $1 \times P$ for the paragraph that contains the focal word. In this scenario, P corresponds to the number of paragraphs in the corpus. Furthermore, a new weight matrix D with the dimension of $P \times N$ is introduced in this model. This matrix is used to propagate the values in the paragraph ID input vector to the hidden layer similarly to the way it is done for the word embedding vectors with the weight matrix W. Next, the algorithm works similarly to the one in the CBOW model. The input vectors of word embeddings are multiplied by the weight matrix W and the input vector of paragraph ID is multiplied by the weight matrix D which results in a total of $C + 1$ embedding vectors with dimension $1 \times N$. This one extra embedding vector is extracted from matrix D and it corresponds to the embedding vector of the paragraph that contains the focal word. The values in the hidden layer are obtained by taking the average of these vectors which results in a hidden layer vector with the dimension $1 \times N$. The further steps are equivalent to the ones in the CBOW model, thus, the values are propagated forward to the next layers of the network to arrive at the prediction for the next word in the paragraph provided the context. Next, the loss function is used to compare the obtained prediction with the actual focal word. Afterward, the values in the weight matrices are produced by iteratively applying the backpropagation and stochastic gradient descent algorithms until all the training instances go through the network. Following a similar logic as in the case of word embeddings, the PV-DM model tries to map the embeddings of the paragraphs that share their semantic meanings closer to each other in the latent vector space.

[6] Source - Distributed Representations of Sentences and Documents paper by Le & Mikolov (details in the References list)

While training the Doc2Vec model it is important to consider the values which are selected for the hyperparameters to ensure the optimal performance of the model. In this study, two hyperparameters are tuned to arrive at the optimal result namely the window size C and the learning rate at the end of the training (in the last epoch). The usual way of tuning the hyperparameters is to perform the classification while using the embeddings of the paragraphs as the inputs of the model. For this task, labeled data is required. All the instances in the data set can be split into the training and testing sets and the training set can be used for the training of the model for prediction purposes while the testing set can be used as the new unseen data for evaluation purposes. However, for this research, the goal is to obtain the embedding representation of all the summaries so that they can be compared and evaluated. Hence, it is not feasible to leave out a test set as these instances or summaries will not be mapped to the embedding vectors. Instead, this paper utilizes an alternative approach for hyperparameter tuning. This approach implies the selection of the appropriate values for the hyperparameters by judging the general consistency of the Doc2Vec model. This method is outlined in the Python library called Gensim which offers a guide on the implementation of the Doc2Vec model (Řehůřek, 2019). First, a new embedding vector is inferred for each paragraph from the trained model. These new vectors will be considered as new unseen paragraphs. Inferring new vectors can be done by fixing the weight values in the word embeddings matrix and by introducing the "new" paragraph in the paragraph embeddings matrix after which the backpropagation and stochastic gradient descent algorithms are applied to produce the embedding vector representation of the new paragraph. After that, the inferred vectors are compared with the training corpus by calculating the similarities between the inferred vectors and all pre-trained embeddings of the paragraphs. Lastly, the rank of the original paragraph is returned based on self-similarity. In an ideal scenario, if the obtained model is consistent the paragraph will be most similar to itself and hence, the rank for the paragraph itself will be one. After applying this method on all the paragraphs in the model the proportion of the paragraphs that are regarded as most similar to themselves (proportion of the ranks which are equal to one) is computed. This proportion has been calculated by varying the values of two hyperparameters and the combination that resulted in the highest value of the proportion has been selected to configure the optimal Doc2Vec model.

Once the vector representation of the article summaries is obtained these embeddings are saved in the CSV (comma separated values) format so that they can be loaded back to RStudio for

further analysis. The embedding representation of each of the four summaries is stored in four separate data frames with 100 rows corresponding to the summary of each article and 100 columns corresponding to each dimension. The next step comprises the selection of an appropriate distance measure. While there is an availability of many various distance measures such as Manhattan or Euclidean distance it is important to select the most suitable method diligently. The guidance on selecting the most suitable distance measure that would provide a meaningful concept of the proximity between two entries in the academic literature is very limited. Aggarwal, Hinneburg, & Keim (2001) have shown that in a high-dimensional space the relative divergence of the distances to a point of interest is heavily dependent on the L_k metric which is used. The research provided evidence that when dimensionality increases the meaningfulness of the L_k norm decreases faster for the higher values of k. Consequently, it is preferable to choose the lower values of k for the problems which involve high dimensionality. It is logical then to conclude that a distance measure with L_1 norm such as Manhattan distance would perform better than a distance measure with L_2 norm such as Euclidean distance in the high dimensional space.

Following this logic, Aggarwal et. al. (2001) propose a new so-called fractional distance metric that allows the value of k to be a fraction smaller than 1. The empirical evidence demonstrates and proves that this new metric does a better job at retaining the meaningfulness of proximity measures. The formula for calculation of the fractional distance metric $dist_d^f$ with the L_f norm between x and y can be represented in the following manner:

$$dist_d^f(x, y) = \sum_{i=1}^{d} [(x^i - y^i)^f]^{1/f},$$

where f is the value between 0 and 1 and d corresponds to the number of dimensions. This formula will be used to calculate the distances between all the generated extractive summaries (Global summary, K-means summary, and Spherical K-means summary) and the golden abstractive Highlights summary. Once those distances are calculated, they will be compared for all 100 records and the similarities between the summaries will be determined.

Therefore, after the comparison is made it will become clear as to which out of extractive summaries lies closer in the embedding space to the golden standard Highlights summary for all 100 records. Once those proportions are obtained, it is possible to verify whether the differences in proportions are statistically significant or not. This can be accomplished by the

means of the Binomial test for proportions (for the detailed description please refer to the appendix). Here the null hypothesis would be formulated as: the true probability of success is equal to 0.5, and the alternative hypothesis would be formulated as: the true probability of success is not equal to 0.5. In case if the p-value would turn out to be less than 5% we would have enough evidence to reject the null hypothesis at 5% significance level.

Since the distances between all the produced summaries and the golden standard summary are available it is possible to determine the summary which is closest to the golden standard summary and the summary which is situated the furthest from the golden standard summary in the high dimensional space. This way, it is possible to have a qualitative examination of those summaries and give some assumptions on why the distance is so small/large. Hence, these summaries will be extracted and explored further at the end of the analysis.

Chapter 5

Results

The previous chapter described all the methods that were applied in this study. Keeping in mind the main goal of this research that is evaluating the quality of the produced summaries and determining whether summaries produced post-application of clustering have a better quality than the global summary, this chapter will focus on the elaboration of the obtained results. The first section will focus on the results produced by the clustering algorithms and how these results were utilized for the generation of the summaries via the LexRank algorithm. Later the output of the Doc2Vec model will be addressed by explaining the choices that were made for the tuning and obtaining the optimal model for paragraph embeddings. Lastly, the chapter will expand on the evaluation part where the comparison of summaries and their qualitative evaluation will be discussed.

Summaries

A total of three summaries are produced and compared to determine the summary that is the most similar to the golden Highlights summary. The first summary called Global summary has been produced by the LexRank algorithm without application of the clustering meaning that the summarization algorithm assumes that all the input sentences belong to one cluster. Some

of the preprocessing steps that are required in order to apply the LexRank algorithm on each article in the data set were described in the Data chapter. The argument "continuous" has been set to TRUE to ensure that the algorithm utilizes the continuous LexRank meaning that the threshold will be ignored and LexRank scores will be calculated using a weighted graph representation of the sentences. The output of the function is a new data frame of sentences and their corresponding LexRank scores. After that, the top four sentences with the highest LexRank scores have been selected to be included in the final Global summary for each article. These steps have been applied to each article in the data set and the corresponding four-sentence summaries have been obtained.

The second summary called K-means summary has been produced by first applying the clustering technique to the articles in the data set. Before applying the clustering algorithm on each article in the data set, some preprocessing of the data is required and these steps can be found in the Data chapter. Once the input for the algorithm is ready a seed is set to provide reproducibility. The "kmeans" function that was used to partition the sentences requires specification of the number of clusters that are desired, thus, the argument has been set to produce four clusters. Once the results of the algorithm have been produced an indication of which sentence in an article belongs to which cluster is obtained. Consequently, all the sentences in an article were separated according to their cluster belonging and they were stored in four separate data frames with their corresponding sentence IDs. Next, the LexRank algorithm has been applied to each data frame separately in a similar manner as it has been described in the section above for the Global summary. However, in this case, it is possible that for some clusters the algorithm selects only one sentence. Hence, it is not pertinent to apply the LexRank algorithm to just one sentence. In this case, if any cluster contains only one sentence then instead of applying the LexRank this sentence gets automatically selected to be included in the final summary for the corresponding article. Otherwise, the LexRank algorithm is applied to all the sentences in each cluster and their LexRank scores are determined and stored in the new data frames. Following that, a sentence that corresponds to the highest LexRank score from each cluster is selected to be included in the final summary for the related article. Thus, each cluster contributes one sentence to the final summary for each article. This method allows us to account for different topics present in each article since the top-ranked sentence is selected not from the whole article but each of four different clusters. Therefore,

sentences that normally would appear at the bottom of the list are now assigned to the higher LexRank scores and, thus, get higher chance of being included in the final summary.

The third summary is named Spherical K-means summary since it utilizes another clustering approach before applying the LexRank algorithm. All the steps to produce this summary are equivalent to the ones used for the generation of the K-means summary with the only difference being the employment of a spherical k-means algorithm. After applying all the preprocessing steps to the articles (similar to the ones applied for the k-means algorithm) and setting the seed for reproducibility a DTM is used as an input for the spherical k-means algorithm. The function that is used to produce the clusters called "skmeans" accepts some extra arguments that need to be specified. Similarly to the "kmeans" function, the number of desired clusters needs to be specified and thus, it has been set to four. Another parameter named m controls the softness or fuzziness of the partition where the value of one would mean a hard partition and values greater than one produce partitions of increasing softness. The value of this parameter has been set to 1.2 to allow for some fuzziness in the partition. Additionally, two control parameters have been specified as follows. The argument $nruns$ has been set to 5 to perform the fixed-point runs five times and the argument $verbose$ has been set to TRUE to monitor the output on minimization progress while the algorithm is running. After all the sentences in an article are assigned to their corresponding clusters the sentence with the highest LexRank score from each cluster is selected to be a part of the final summary for the article. Similarly to the K-means summary in case if any cluster contains only one sentence this sentence gets automatically selected to be a part of the final summary for the corresponding article.

Doc2Vec Results

Once the summaries are loaded and stored in a data frame the preprocessing steps are applied to prepare the data for the Doc2Vec model. These cleaning and preprocessing steps were outlined and discussed in the Data chapter. Doc2Vec model requires that the model is trained on the tagged data. Hence, to train the model each article has to be first associated with a tag or number for the entire training corpus. Once each article is assigned a tag this can be used as an input of the model. Some of the parameters had to be specified prior to the training of the model. Thus, the learning rate at the beginning of the training has been set to 0.01, the number of training cycles or epochs has been set to 100, and the vector size has been set to 100 to obtain the vector representation of the summaries in 100 latent dimensions. Additionally, a random

seed has been specified for reproducibility. To identify the most optimal Doc2Vec model some choices have to be made with regard to the selection of the appropriate values for the hyperparameters. The values that were tested for the hyperparameter window size are 2,3, and 4 and the values that were tried for the hyperparameter minimum alpha (which corresponds to the learning rate by the end of the training) are 0.001, 0.0001, and 0.00001. These three options for each hyperparameter resulted in a total of nine combinations. The percentages of times when the model generated the embedding vectors which ranked the summary as the most similar to itself for each of the four summaries are illustrated in Table 2. It can be seen that the range of the percentages differs across the four models for each of the summaries. The highest value (97%) has been obtained for combinations number 7, 8, and 9 for the Spherical K-means summary and the lowest value (42%) has been obtained for combination 3 for K-means summary. In general, all combinations for Highlights and Spherical K-means summaries resulted in the corresponding models ranking a summary as the most similar to itself between 94% and 96% as well as 96% and 97% of the times respectively. However, the results for the Global and K-means summaries demonstrate relatively low percentages meaning that these models are less robust with respect to the fluctuations in the data. Thereby, combinations 1 and 7 with the corresponding 96% and 97% were selected to build the optimal PV-DM model for the Highlights and Spherical K-means summaries respectively. Moreover, combinations 7 and 9 with the accordant 79% and 66% were chosen to produce the final PV-DM model for the Global and K-means summaries respectively. Nevertheless, the values in Table 2 indicate that other configurations could also be used for the generation of the optimal PV-DM models for the Highlights and Spherical K-means summaries since the model performance would not decline sharply. This is not the case for the Global and K-means summaries, however, because the range of the percentage values is much larger for them as compared to the other summaries. Once the optimal models have been generated for each summary, the obtained embeddings matrices were stored in four separate data frames. Each data frame consists of the 100 rows corresponding to the 100 article summaries and of the 100 columns that correspond to the 100 latent dimensions. The values in the data frames correspond to the embeddings of the 100 summaries for each dimension.

Table 2: These are 9 combinations of hyperparameters for tuning of PV-DM model of Doc2Vec. The values correspond to the percentages of times when the model generated the embedding vectors which ranked the summary as the most similar to itself for each of the four summaries. Low percentages indicate that these models are less robust with respect to the fluctuations in the data.

Combination	Window Size	Minimum Alpha	Highlights - % of times a sentence is ranked as similar to itself	Global - % of times a sentence is ranked as similar to itself	K-means - % of times a sentence is ranked as similar to itself	Spherical - % of times a sentence is ranked as similar to itself
1	2	0.001	96%	60%	44%	96%
2	3	0.001	96%	60%	45%	96%
3	4	0.001	96%	60%	42%	96%
4	2	0.0001	94%	78%	59%	96%
5	3	0.0001	94%	78%	60%	96%
6	4	0.0001	94%	78%	60%	96%
7	2	0.00001	94%	79%	62%	97%
8	3	0.00001	94%	79%	64%	97%
9	4	0.00001	94%	79%	66%	97%

Quantitative Evaluation

Once the embedding vector representations of the article summaries are obtained it becomes possible to calculate the distances between the summaries in the embedding space. This would indicate how close the produced summaries lie to the golden standard Highlights summary. To do this, using an appropriate fractional distance measure the distances between the Global and Highlights summaries, the K-means and Highlights summaries, and finally the Spherical K-means and Highlights summaries are calculated. This has been done for each and every row of the data frame, hence, the result comprises the new columns in the data frame that indicate the numerical values of the distances between the corresponding summaries. As soon as these numbers are calculated, the next step includes the comparison of those distances.

Before arriving at the conclusion on which summary out of the generated summaries is the closest to the Highlights summary, three statements have been formulated to assess the status of the distances between the related summaries. The first statement is formulated as follows: *"The distance between the Highlights summary and the Global summary is less than the distance between the Highlights summary and K-means summary"*. If this statement turns out to be correct this would indicate that the Global summary is better than the K-means summary since it lies closer to the golden standard summary in the embedding space. To check this

statement the distances for all hundred summaries have been compared which resulted in rejecting the first statement. The analysis showed that for all hundred rows the distances between the Highlights summary and the K-means summary are smaller than the distances between the Highlights summary and the Global summary. This indicates the better quality of the K-means summary over the Global summary.

The second statement is constructed in the following manner: *"The distance between the Highlights summary and the Global summary is less than the distance between the Highlights summary and the Spherical K-means summary"*. Similarly to the previous example, the correctness of this statement would indicate that the Global summary is of a better quality as compared to the Spherical K-means summary. The comparison of the distance values for all hundred summaries showed that the Spherical K-means summary lies closer to the Highlights summary than the Global summary. Thus, the second statement is also rejected meaning that the Spherical K-means summary has a better quality than the Global summary with respect to the Highlights summary.

The last statement compares the distances between the K-means summary and the Spherical K-means summary since they both turned out to be of a better quality than the Global summary and it is formulated as follows: *"The distance between the Highlights summary and the K-means summary is less than the distance between the Highlights and the Spherical K-means summary"*. The correctness of this statement would indicate that the K-means summary has a higher quality as compared to the Spherical K-means summary. In this case, the comparison resulted in rejecting the third statement for 76 out of 100 rows. This suggests that in 76 out of 100 times the Spherical K-means summary has a better quality than the K-means summary as it lies closer to the Highlights summary.

Now that the proportion is obtained a Binomial test for proportions can be used to verify whether this difference in proportions is statistically significant. Here, let us define 76 as the number of successes and 100 would be the number of total trials n. Assuming that the probability of success (p) for any single observation is 50% the null hypothesis is defined as *"The true probability of success is 0.5"* and the alternative hypothesis is defined as *"The true probability of success is not 0.5"*. The obtained results demonstrate that with the 95% confidence interval being between 0.66 and 0.84 the p-value turned out to be equal to 1.81e-07

which is a vastly small number. This evidence gives us the confidence to reject the null hypothesis with the significance level of 5%. These findings prove that the obtained results are significantly different from the null hypothesis.

Qualitative Evaluation

Since the distances between the summaries are calculated it is possible to check which article summaries are the closest to the golden standard summary and which are the furthest away. This allows extracting and examining these summaries to get more insights into why they lie closer or further away from the golden standard summary. Thus, first, the summaries (Global, K-means, Spherical K-means) that are located the closest according to their distances to the golden Highlights summary will be reviewed and examined and later the summaries that are located the furthest away will be evaluated.

Due to the original articles being very lengthy they are presented in the appendix of this paper. Additionally, Table 3 contains the top three summaries that are the closest to the Highlights summary. The qualitative evaluation of the articles alongside their summaries discussed in this paper reflects the views of an author only. First, let us look at the Global summary that is the closest to the Highlights summary. This is a second article in the data set and it mainly talks about different ways of celebrating Pi day and how it all started. The corresponding Highlights and Global summaries of this article can be found in Table 3. It is important to keep in mind that the Highlights summary is abstractive meaning that the new sentences were composed for this summary, whereas, the Global summary is extractive meaning that it contains the selected sentences from the original article. The Highlights summary provides an overview of what is Pi and how people started celebrating it whereas, the Global summary contains four very short sentences describing an indication about the Pi day celebration. The sentences that ended up in the Global summary for this article provide a hint on the different topics discussed in the original article, however, they do not provide much information about these topics. In this sense, an algorithm managed to grasp important parts of the article without application of clustering, hence, without picking the important sentences from different clusters.

Moving on to the K-means summary, an article for which the K-means summary turned out to be the closest to the Highlights summary is article number 20 and it can be found in the appendix. To sum up, this article primarily talks about the rescue of boys and men who have

been held captive in one of the Islamic religious schools in Pakistan. The Highlights summary of this article that can be found in Table 3, summed up all the important points of an original article by rephrasing some of the sentences. On the other hand, the K-means summary for this article combined the most important sentences from four different clusters of sentences. This summary managed to include the information not only about the main news but it also talks about the fact that the school was a drug rehab clinic during the day time and that the operation was successful. It is interesting to note that it also contains a sentence about one woman being willing to pay the police to keep her troublesome child. The Highlights summary does not mention this point, however, it grabs the attention of a reader in the original article. Overall, an algorithm managed to grasp the most important topics of the original article for the K-means summary.

As for the Spherical K-means summary, article number 38 resulted in the summary that is the closest to the Highlights summary. This article talks about the horrible consequences of an earthquake that happened in Haiti a week before the biggest Carnival of the nation. Many singers lost their homes and musical instruments in that earthquake and they will never be able to play again. However, the Haitians stay strong and have faith in rebuilding their future. The article itself is quite lengthy and even the Highlights summary did not mention all the topics that were covered in the original article, though it touched upon the main idea. The Spherical K-means summary of this article does not include a sentence about the earthquake specifically. However, it includes the quote of a nurse who lost her home and her belongings in the earthquake. Thus, overall, it reflects quite clearly on the main idea of an article. The selection of sentences for this summary is quite interesting since it includes a sentence that talks about how a singer is set for the rest of the year if he performs well in the Carnival. This topic was not present in the Highlights summary but it grasps the attention of a reader when reading the original article.

Therefore, overall, these summaries managed to grasp the most important topics of the original articles and convey the information in a clear manner. When reading the summaries one can notice that the application of clustering for the K-means and Spherical K-means summaries helped to identify and select sentences that talk about different topics within the articles.

Table 3: The closest summaries to the Highlights summary

Article Number	Highlights Summary	Global Summary	K-means Summary	Spherical K-means Summary
2	Math geeks and others celebrate Pi Day every March 14. Pi, or roughly 3.14, is the ratio of circumference to diameter of a circle. The Pi Day holiday idea started at the Exploratorium museum in San Francisco. Albert Einstein was also born on March 14.	That's why March 14 - 3-14 - is Pi Day. Where Pi Day began. Even more pi. How do you celebrate Pi Day?		
20	Captive boys and men were rescued from an Islamic religious school in Pakistan. They were reunited with their families this week. The facility was a school and drug rehab clinic. Authorities say they're searching for the owners; three others arrested at the facility.		(CNN) -- The 54 men and 14 boys rescued after being found chained this week at an Islamic religious school in Pakistan have been reunited with their families or placed in shelters, authorities said. The school, Al-Arabiya Aloom Jamia Masjid Zikirya, which also was a drug rehab clinic, is in Sohrab Goth, a suburb of Gadap in Karachi. "The operation was successful, and we plan on continuing our work to ensure that places like this are shut down," Marwat said. One woman told a local television station that she was willing to pay the police to keep her troublesome child.	
38	Less than week ahead of Haiti's Carnival celebration, revelry replaced with mourning. Haitians have celebrated Carnival through dictatorships, military coups and bloodshed. "I don't even remember when it is," Haiti official says about upcoming three-day festival.			They will not be able to play again."We're living in a city that's like a cemetery," said Ronide Baduel, a nurse who lost her home and all her belongings in the quake. "If you have a good showing at Carnival, you're set for the rest of the year," Martino said. Carnival, he said, was Haiti's musical showcase.

Moving on to the evaluation of the produced summaries that are the furthest away from the golden standard Highlights summary, the original articles for these summaries can be found in

the appendix of this paper and the summaries are presented in Table 4. Starting with the Global summary, an article for which an algorithm produced the furthest Global summary from the Highlights summary is article number 66. This article describes an interview with Stella McCartney where she talks about her work and what influence her famous parents had on her fashion career. She also emphasized the importance of sustainability in the fashion industry by highlighting the fact that she never works with fur or leather materials. The Highlights summary managed to capture all the main points of the original article by mentioning her career, her belief in a more sustainable fashion industry, as well as, the fact that she won a prestigious award. The Global summary of the same article, however, consists of the selection of sentences that mainly talk only about sustainability in fashion and the fact that she does not use leather material. Whereas it does not include sentences describing the influence of her famous parents or her winning the award which are quite important topics of the article. Keeping in mind that this summary has been produced by applying a summarization algorithm to the whole article (and not to separate clusters of sentences) this may explain why sentences with other topics were not selected to be included in the final summary.

An article for which the distance between the produced K-means summary and the Highlights summary is the largest as compared to the remaining article summaries is article number 18. This article highlights the news that the leaders of America, Canada, and Mexico came to an agreement to streamline border controls to provide the free movement of people through the trusted program for travellers. Furthermore, the article contains the quotes of the corresponding leaders and other important people who are involved in the discussions that emphasize their opinions regarding the discussed topics. Judging the summaries of this article, it is worth mentioning that the Highlights summary for this article contains three rather short sentences, and while it mentions the agreement between the three parties it does not elaborate enough on the details. On the other hand, the K-means summary for this article contains the selection of four sentences that mention all the important topics of the article. Apart from the main point that was mentioned earlier, it also mentions Obama's (president of the USA at the time) commitment to the immigration reform, the discussion on the trade agreement with Pacific nations, as well as Latino special interest group's criticism of the administration for its aggressive deportation policy. Given that these were the main highlights of an article, overall, the produced K-means summary can be described as of good quality since it managed to capture the substantial points of the original article. There is, however, a logical explanation as

to why the distance between this K-means summary and the Highlights summary is so large. Since for this article, the Highlights summary is not explaining the subject of an article in enough detail and provided that the Highlights summary contains only three very short sentences it makes sense that the distance between the two summaries is not that small.

Lastly, article number 71 turned out to be the one for which the distance between the produced Spherical K-means summary and the Highlights summary is the largest. This article provides a review of some features for the new (at a time) Apple products - iPad Mini and iPad Air. An author compares these new products to some previous and existing Apple products in terms of technology and design. An overview of the comparison with some other brands is also provided in an article. Looking at the Highlights summary for this article one can notice that it briefly summarizes some general remarks about the new products and it contains one sentence for each of the new iPads describing their qualities. The Spherical K-means summary for the same article, however, contains two sentences about the Retina display of the Apple products and the remaining sentences comment on Apple's pricing strategy and the need for an improvement of the battery life. Hence, it becomes clear that these two summaries have relatively different content and this could be the reason for the higher distance between the two summaries.

This qualitative framework provides an opportunity for the detailed evaluation of the produced summaries and it gives some insights to determine why some distances between the summaries are smaller and others are larger. The downfall of this evaluation method, nevertheless, is the fact that it is a time-consuming method and normally the evaluation of a few human readers would be required to arrive at robust conclusions.

Table 4: The furthest summaries from the Highlights summary

Article Number	Highlights Summary	Global Summary	K-means Summary	Spherical K-means summary
66	McCartney talks about the influence of her famous parents on her fashion career. Says she doesn't see leather as luxury, as it's mass produced. Wins prestigious Women's Leadership Award from Lincoln Center Corporate Fund. Believes it's not sustainable to kill so many animals for shoes and handbags.	CNN: Do you feel a lot of pressure from the industry because you chose to go down the "sustainable fashion" road? I've had people say to me: "You'll never sell handbags, you don't work with leather and leather is luxury." Ninety percent of the people who come to		

		my stores have no idea I don't work with leather. I'm learning as I get older, that you don't have to try to fight everything from a man's place.		
18	Three nations will streamline border controls. The three leaders are called the "Three Amigos". The Keystone XL pipeline is a major issue.		President Barack Obama, Canadian Prime Minister Stephen Harper and Mexican President Enrique Pena Nieto agreed to streamline border controls to facilitate the movement of people through the establishment of a trusted traveler program. They also spent a great deal of time during the North American Leaders' Summit discussing efforts to broker a new trade agreement with Pacific nations, they said in a joint news conference after the summit in Toluca, Mexico. Latino special interest groups, a core Democratic Party constituency, have criticized the administration for its aggressive deportation policy as it struggles to find a long-term solution to the immigration issue. Senior administration officials told reporters last week that Obama remains committed to comprehensive immigration reform that includes a pathway to citizenship for undocumented workers.	
71	The new iPads from Apple don't hold a lot of surprises, but the regular improvements are solid. The renamed iPad Air is thinner, faster and lighter than the previous generation. With a new processor and better screen, iPad Mini catches up to the full-size iPad.			A number of Apple products already have the Retina Display: the iPhone 4S and up, the fourth generation iPod Touch, the 10-inch iPad and some MacBook Pros. Like the iPad Air, the iPad Mini with Retina has the A7 chip inside. Other companies might throw in price drops, but Apple is comfortable in its spot at the high-end of the consumer market. Faster processors gobble up more battery power,

				so maintaining 10 hours of battery life from an old tablet to a new tablet probably requires battery improvements.

Chapter 6

Conclusion

The goal of the research conducted in this paper was to assess the quality of the produced summaries and to compare them against a benchmark that is a golden standard Highlights summary to determine which summary has a better quality. Therefore, the following research question has been formulated:

Does the application of clustering algorithms help to improve the quality of summary as compared to the global summary?

To address this research objective, the distances between the produced summaries and the Highlights summary were calculated and compared. Two summaries that were produced after application of the K-means clustering and the Spherical K-means clustering algorithms were compared to the Global summary that was produced without application of the clustering against the Highlights summary that was provided with the data. Once again, it is important to keep in mind that the produced summaries are of the extractive form, whereas, the Highlights summary is of abstractive form. The quantitative analysis demonstrated that the distance between the K-means summary and the Highlights summary for all 100 observations is smaller than the distance between the Global summary and the Highlights summary. This provides evidence to conclude that the generated K-means summary is closer to the golden standard Highlights summary and thus, overall it has better quality than the Global summary. At the same time, the distance between the Spherical K-means summary and the Highlights summary for all 100 observations turned out to be smaller than the distance between the Global summary and the Highlights summary. This suggests that the Spherical K-means summary is closer to the Highlights summary and hence, it is of better quality than the Global summary. Provided this information, it is possible to conclude that overall given 100 articles and their corresponding summaries, the summaries that were produced after application of the clustering

algorithms demonstrated a better quality when compared to the global summary that was produced without application of the clustering algorithm by using the golden standard Highlights summary as a benchmark.

To comment on the relative performance of two different clustering algorithms that were applied in this research, the study concluded that with the significance level of 5% for 76 out of 100 cases the Spherical K-means summary is significantly closer to the Highlights summary as opposed to the K-means summary. This information indicates that roughly 76% of the time the Spherical K-means summary has better quality when compared to the K-means summary. Therefore, the Spherical K-means clustering algorithm managed to create a partition of sentences in the articles such that the summarization algorithm could grasp the most important sentences from each cluster. Even though the Spherical K-means clustering algorithm produced such useful partitions, the K-means clustering algorithm still demonstrated significant results as the K-means summary turned out to have a better quality than the Global summary.

When it comes to the qualitative evaluation of the summaries, the analysis of the summaries that are located the closest and the furthest away from the Highlights summaries showed that generally there is a benefit of applying the clustering algorithms before producing the summaries because it helps in identifying the most important sentences that are related to different topics within the corresponding articles. For some articles, however, the distances between the given produced summary and the Highlights summary turned out to be very large. This can be explained by the fact that for those articles the provided Highlights summaries did not include all the important details of the articles or they contained the same information but in a rephrased form such that it led to the larger distances between the summaries. Hence, a more careful human reader evaluation may have resulted in different findings.

The abovementioned findings suggest that it is important to consider the various possible topics within a given article when producing the summary. As demonstrated in this paper, it can be achieved by introducing the nonlinearity component where the main focus is not only on the centrality but also on the other topics in an article. In this method, the sentences that would normally get a low LexRank score and appear at the bottom of the list are assigned a higher LexRank scores which increases their chance of being a part of the summary. These would be

the sentences that are important to each of the obtained clusters but are not centered. This way, centrality alone does not dictate the importance of the sentences.

The following paragraphs will address the limitations of this paper and will provide suggestions for future research. Thus, firstly, it is worth mentioning that this quantitative research was conducted with only a hundred observations due to the time and computational capacity constraints. The accommodation of more data would help to arrive at more robust results, however, it is important to keep in mind that conducting analysis on more data would also require more time and resources. Another idea would be to conduct a similar analysis on a different data set to be able to compare the results and identify whether the findings match. The application of the methods used in the analysis of this paper when applied to other types of text data could result in different conclusions.

Moreover, due to the selection and application of the specific techniques to partition sentences in the articles the obtained results might differ when using other methods. There are some other methods that could help in identifying different topics and splitting the sentences within the articles. For instance, a widely popular topic modelling approach called Latent Dirichlet Allocation (LDA) could be used to classify the documents and identify a number of different topics within those documents where the sets of observations are explained by unobserved groups which provide an indication of why some parts of the documents share certain similarities. Different approaches to sentence partitioning may potentially result in completely different sets of sentences being selected to be a part of the final summary. Additionally, as this paper utilized a continuous LexRank method for the generation of summaries, the use of other automatic summarization techniques may be accommodated to verify the stability of the results.

Another limitation of this research is the fact that the sentences were split into only four clusters for the generation of the post-clustering summaries such as K-means and Spherical K-means summaries. A different number of clusters was tried out, however, it has been eventually decided to use only four to save time for the evaluation of the summaries. With only four clusters the analysis has resulted in a comparison of four summaries and if more clusters would be used to generate the summaries the evaluation and comparison of so many summaries would not be feasible for this paper. However, it would be a good idea to try out and compare the

quality of summaries that get generated after splitting sentences into different numbers of clusters. Judging by the quality of these summaries one can also comment on the optimal number of clusters for a given clustering algorithm and a given data set.

Moving on, another limitation in this research is the unavailability of a human reader who would read and compare the quality of the produced summaries. Ideally, an expert reader would be able to judge the summaries and provide some comments on their quality to understand the pitfalls of the method. This would help to take a closer look at the summaries and determine the small aspects that could have affected the quality of a summary. Due to the unavailability of a human reader, this paper utilized an alternative approach to the evaluation of the summaries by obtaining a vector embedding representation of each summary and by calculating the distances between the summaries in a latent vector space. These results, however, may differ from the objective feedback that could be obtained in a presence of an expert human reader. Hence, it would be beneficial to have a human reader evaluate the quality of the summaries in future research.

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001, January). On the surprising behavior of distance metrics in high dimensional space. In *International conference on database theory* (pp. 420-434). Springer, Berlin, Heidelberg.

Alsabti, K., Ranka, S., & Singh, V. (1997). An efficient k-means clustering algorithm.

Bergamaschi, S., Guerra, F., & Leiba, B. (2010). Guest editors' introduction: information overload. *IEEE Internet Computing*, *14*(6), 10-13.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual web search engine. Computer Networks, 30, 107–117.

Buchta, C., Kober, M., Feinerer, I., & Hornik, K. (2012). Spherical k-means clustering. *Journal of Statistical Software*, *50*(10), 1-22.

Chapman, C., & Feit, E. M. (2015). *R for marketing research and analytics* (p. 195e223). New York, NY: Springer.

Dhillon, I. S., & Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine learning*, *42*(1-2), 143-175.

Dhillon, I. S., Guan, Y., & Kogan, J. (2002, April). Refining clusters in high-dimensional text data. In *Proceedings of the workshop on clustering high dimensional data and its applications at the second SIAM international conference on data mining* (pp. 71-82). SIAM.

Erkan, G., & Radev, D. R. (2004). Lexrank: Graph-based lexical centrality as salience in text summarization. *Journal of artificial intelligence research*, *22*, 457-479.

Gambhir, M., & Gupta, V. (2017). Recent automatic text summarization techniques: a survey. *Artificial Intelligence Review*, *47*(1), 1-66.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT press.

Harabagiu, S., & Lacatusu, F. (2005, August). Topic themes for multi-document summarization. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 202-209).

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100-108.

Hennig, L. (2009, September). Topic-based multi-document summarization with probabilistic latent semantic analysis. In *Proceedings of the International Conference RANLP-2009* (pp. 144-149).

Ko, Y., & Seo, J. (2008). An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. *Pattern Recognition Letters*, *29*(9), 1366-1371.

Kwartler, T. (2017). *Text mining in practice with R*. John Wiley & Sons.

Le, Q., & Mikolov, T. (2014, January). Distributed representations of sentences and documents. In *International conference on machine learning* (pp. 1188-1196).

Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165.

Mann, W. C., & Thompson, S. A. (1988). Rhetorical structure theory: Toward a functional theory of text organization. *Text*, *8*(3), 243-281.

Mihalcea, R., & Tarau, P. (2004, July). Textrank: Bringing order into text. In *Proceedings of the 2004 conference on empirical methods in natural language processing* (pp. 404-411).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Plaza, L., Díaz, A., & Gervás, P. (2010). Automatic summarization of news using WordNet concept graphs. *IADIS International Journal on Computer Science and Information Systems*, *45*(57).

Porter, M. F. (2001). Snowball: A language for stemming algorithms.

Řehůřek, R. (2019). Gensim: Doc2vec model. https://radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec_lee.html

Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of documentation*.

Salton, G., Singhal, A., Mitra, M., & Buckley, C. (1997). Automatic text structuring and summarization. *Information processing & management*, *33*(2), 193-207.

Sethi, P., Sonawane, S., Khanwalker, S., & Keskar, R. B. (2017, December). Automatic text summarization of news articles. In *2017 International Conference on Big Data, IoT and Data Science (BID)* (pp. 23-29). IEEE.

Tas, O., & Kiyani, F. (2007). A survey automatic text summarization. *PressAcademia Procedia*, *5*(1), 205-213.

Wong, K. F., Wu, M., & Li, W. (2008, August). Extractive summarization using supervised and semi-supervised learning. In *Proceedings of the 22nd international conference on computational linguistics (Coling 2008)* (pp. 985-992).

Xu, R., & Wunsch, D. (2005). Survey of clustering algorithms. *IEEE Transactions on neural networks*, *16*(3), 645-678.

Zhong, S. (2005, July). Efficient online spherical k-means clustering. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.* (Vol. 5, pp. 3180-3185). IEEE.

Appendix

Binomial Test for Proportions

The Binomial test for proportions assumes that all observations fall into only two categories namely ‘success’ and ‘failure’ and the probability of success for each observation is identical. For the sake of example, let us assume that the number of observations for which the distance between the Highlights summary and the Spherical K-means summary is smaller than the distance between the Highlights summary and the K-means summary is X. Then X would be considered as a number of ‘successes’ and it will be following the binomial distribution with the parameter n and p. Here the parameter n corresponds to the total number of observations (in our case it is equal to 100) and the parameter p would be the probability of success for any single observation. The question to ask in this case would be as follows. Is the observed value X significantly different from the equal representation which would correspond to the p-value of 50%? The function “binom.test” is used to verify the likelihood of randomly observing X cases out of 100 in one direction if the true likelihood is 50%. The null hypothesis, in this case, would be that the true probability of success is equal to 0.5 and the alternative hypothesis would be that the true probability of success is not equal to 0.5. Using the significance level α of 5% if the p-value turns out to be less than 5% the null hypothesis will be rejected. This would support the idea that the results are significantly different from the null hypothesis. The p-value which is calculated by the function corresponds to the probability of the given test statistics taking the value as extreme or even more extreme than the actually observed one assuming that the null hypothesis is true. Hence, the small p-value would indicate more evidence against the null hypothesis.

Article number 2

"March 14 is my favorite day to be a nerd. Across the country, math geeks in museums, schools, private groups and elsewhere gather to celebrate the number pi, approximately 3.14. That's why March 14 - 3-14 - is Pi Day. What's more, Albert Einstein was born on this day. A quick refresher: Pi is defined as the distance around a perfect circle, or the circumference, divided by the distance across it, or the diameter. It is also involved in calculating the area of a circle, the volume of a sphere, and many other mathematical formulas you might need in the sciences. Throughout history, people have been captivated by this number because there is no way to calculate it exactly by a simple division on your calculator. What's more, its digits go on infinitely, without any pattern in the numbers. 3.1415926535897932 ... etc. Even that many digits are more than most people would need for everyday use, but some folks have been

inspired to memorize thousands of digits of pi, or even use the digits to create poetry or music. On Pi Day, one number 'reeks of mystery'. Math may be scary, but pi is not - as evidenced by the widespread revelry on Pi Day. One might even say - gasp! - it's cool to like pi these days. Even the House of Representatives supported the designation of March 14 as National Pi Day in 2009. In countries where the day is written before the month, Friday is 14-3, which looks less like pi. "And so Pi Day is an acquired taste," mathematician Jonathan Borwein, at the University of Newcastle in Australia, said in an e-mail. Conveniently, "pi" sounds like "pie," and pies are round. You could celebrate Pi Day in a casual way by grabbing a slice of pastry, or pizza. If you're in enrolled in school, your math class or math department might be doing something special already. But if you happen to live in a particularly pi-happy place, you might be able to take part in some larger-scale, pi-inspired activities. Where Pi Day began. If you want to go where the day is said to be "invented," look no further than San Francisco's Exploratorium. Larry Shaw, who worked in the electronics group at the museum, began the tradition in 1988. Last year was Pi Day's 25th anniversary there. Pi Day began as a small gathering with mostly museum staff. Now it's a public pi extravaganza featuring a "Pi procession," whose attendees get a number - 0 to 9 - and line up in the order of pi's digits: 3.14159265 ... you get the idea. The parade ends at the "pi shrine" - a pi symbol with digits spiraling around it embedded in the sidewalk, which was unveiled last year. For those who can't attend in person, the Exploratorium has a Second Life Pi Day event that includes "irrational exhibits, fireworks, cheerleaders, music, and dancing." The museum also lists a bunch of educational activities to teach about the concept of pi. On Pi Day, is 'pi' under attack? Where Einstein lived. On the opposite coast, the leafy university town where Albert Einstein spent the last 22 years of his life is showing community-wide exuberance for pi. Princeton, New Jersey, kicks off Pi Day weekend on Thursday night with a reading by physicist Charles Adler, then heads into a full day of activities on Friday, including a walking tour of Einstein's neighborhood and a pizza pie-making contest. The pie-eating contest takes place at McCaffrey's supermarket, while an Einstein look-alike competition will match mustaches and wild gray hair at the Princeton Public Library. Pi fans who have been spending the last year memorizing digits can show off and compete at the library, where the winner among 7- to 13-year-olds can take home a cool pi-hundred (That is, $314.15). The Historical Society of Princeton will have an Einstein birthday party. Tetsuya Miyamoto, inventor of the KENKEN puzzle, will speak at the library as well. Here are 10,000 digits of pi for you to memorize. The "brainiac town" residents "love this event because it's a way for them to celebrate how quirky they are," said Mimi Omiecinski, owner of the Princeton Tour Company, who started Princeton Pi Day in 2009. "A lot of them get super into it." Last year about 9,000 people participated, she said. Along with her fascination with Albert Einstein, Omiecinski was inspired to launch a town-wide Pi Day after she heard that the Princeton University mathematics department celebrates March 14 with pie-eating and pi-reciting (As a Princeton student, I got second place for most digits in 2005 and 2006). Even more pi. Chicago is getting into the pi business too. Lots of restaurants and bakeries are offering Pi Day specials. The Illinois Science Council and Fleet Feet Sports are hosting a 3.14-mile walk/run Friday night, with discounts for anyone named Albert, Alberta or Albertina. Philly.com highlights two options for satisfying your pie cravings in the City of Brotherly Love. Bostonians can head to Massachusetts Institute of Technology at Pi Time (3:14 p.m.) for pi-themed activities such as "Throw Pie at Your Best Friend on High-Speed Camera." The Museum of Science in Boston has educational Pi Day events, and the Seattle Children's Museum will celebrate too. Even the Salvador Dali Museum in St. Petersburg, Florida, will celebrate the day, as "Dali loved the irrational numbers Pi and Phi, often using them and other mathematical principles in his art," according to the museum. If you live in the area, check out their schedule of math-inspired films and tours throughout the day. There are plenty of online resources too, such as piday.org. Outside of the physical

classroom, Pi Day will be celebrated online through Google's virtual classroom project. David Blatner, author of the comprehensive book "The Joy of Pi," is hosting a Pi Day competition in which students from three classrooms will square off to see who can recite the most digits of pi from memory. How did Pi Day become such a big thing? Blatner says that Pi Day has become a hit for the same reason the new "Cosmos" TV show is getting so much attention. "People all around the world are hungry to make science and math fun and interesting," he said in an e-mail. "We know math and science is important, we know that it's fascinating, but we often don't know how to make it fun and interesting. Pi Day gives us a great excuse to throw away our fear of math and say 'Hey, it IS kind of neat!' " If you agree, just wait until 3/14/15 - or as one popular Facebook group calls it, "The Only Pi Day of Our Lives." That's because pi to four digits after the decimal is 3.1415, and we're unlikely to survive until 2115 to see that second instance of pi perfection. So get ready next year to take a picture of your digital clock on 3/14/15 at 9:26:53 a.m. That'll be worth more than a thousand digits. Follow Elizabeth Landau on Twitter at @lizlandau. How do you celebrate Pi Day? Tell us in the comments."

Article number 20

"(CNN) -- The 54 men and 14 boys rescued after being found chained this week at an Islamic religious school in Pakistan have been reunited with their families or placed in shelters, authorities said. The group was discovered in an underground room with heavy chains linking them together. The school, Al-Arabiya Aloom Jamia Masjid Zikirya, which also was a drug rehab clinic, is in Sohrab Goth, a suburb of Gadap in Karachi. All 14 boys were returned to their families, senior police official Ahsanullah Marwat told CNN. Of the adults, 47 had been released to their families, and seven were handed over to a shelter for the homeless, he said. Three people who worked at the facility were arrested, but the four men who ran the place were still at large, Marwat said. Officials said the facility was part madrassa and part drug-rehab facility, and the captives were chained at night apparently to prevent their escape. "The operation was successful, and we plan on continuing our work to ensure that places like this are shut down, " Marwat said. Many of the captives told police their families sent them there because they were recovering drug addicts. During the day, they worked and did religious studies. But the future of the rescued children was unclear. One woman told a local television station that she was willing to pay the police to keep her troublesome child. She said she would rather have the facility remain open, regardless of how it treated the children. Many others, however, said they were in shock and disbelief over the allegations. One man complained he was deep in debt after paying the school a large amount of money to board his son."

Article number 38

"Port-au-Prince, Haiti (CNN) -- In the central plaza, there was once an orgy of music, street dancing and revelry unmatched by any other nation in the Americas, Haitians say. But where there was joy now sits a vast settlement of people left without loved ones, without homes, without life's belongings. Haitians have celebrated Carnival through dictatorships, military coups and bloodshed. Popular belief was that if a government failed to deliver on Carnival, Haiti's equivalent of Mardi Gras, it was sure to fall, said Marie Laurence Lassegue, Haiti's minister of culture and information. But this year, the three-day festival has been canceled, another indication of the enormity of the earthquake's devastation. Musicians fell silent, seamstresses stopped sewing costumes and ghostly skeletons of unfinished floats lay scattered on the outskirts of Port-au-Prince. A month after the devastating January 12 earthquake, the Champs de Mars plaza is home to the capital's displaced, where thousands of people have eked

out a tiny space in which to survive. Full coverage of the earthquake's aftermath. "This is the first time Carnival is not happening," said Roberto Martino, lead singer of popular Kompa band T-Vice. "I don't even think about music anymore." Less than a week ahead of Carnival's start on Sunday, revelry is replaced with mourning. The nation's foremost concert producer, Charles Jubert, died. So did members of four bands who were practicing inside a studio that collapsed. Other musicians lost legs, arms and hands. They will not be able to play again. "I don't think we have time to think of Carnival, " Lassegue said. "Maybe when we are finished crying." Carnival's three days of deliverance and celebration has tremendous importance in the lives of Haitians, Lassegue said. "But this year? I don't even remember when it is." Instead, the displaced are planning days of prayer. "We're living in a city that's like a cemetery," said Ronide Baduel, a nurse who lost her home and all her belongings in the quake. Her brother died and suddenly, she found herself far from her middle-class existence, relegated to a makeshift tent and burlap bags she uses for pillows at night. "I had four good walls around me. Now I have four sheets," she said. She goes to work with a big, black faux-leather purse containing toothpaste, soap and a change of underwear. There, she can bathe properly. "We don't know how many days, how many months, how long we will be this way," she said. "I am always stressed. It's like living in a jungle. How can I dance at Carnival?" Baduel and her tent community neighbors said the money that would have been spent on Carnival ought to be used to build housing. Nearby, the 44 members of Relax Band, who normally would be revving up their street performances in the days before Carnival, worried about their next meal. They played the Sunday before the earthquake, marching through the streets, getting ready for the big performance. Now, everything was gone,- including all their instruments that were crushed when band coordinator Ernst Beauvais' house collapsed. A small stage emblazoned with the red and white logo of Relax Band now harbors a massive water bladder tank dropped off by an aid group and a few mattresses for slumber under the stars. "It is one of the greatest tragedies to befall our country," Beauvais said, pointing to the rubble of his house. He said it was the street band's 30th anniversary; the musicians were looking forward to showcasing their new song. Almost every band in Haiti debuts new pieces at Carnival. On the outskirts of town, the skeletons of three floats sit like ghosts, reminders of what might have come next week. One of the floats belongs to T-Vice. Bandmates Roberto Martino and Eddy Viau would have been practicing with the rest of the band for their Carnival performance --- it's an honor to win top prize. "If you have a good showing at Carnival, you're set for the rest of the year, " Martino said. This year, the band had planned a soccer-themed show with a song called "The End of the Match. "Instead it recently released "Nou Pap Lage" (We Won't Give Up"), dedicated to the victims of the earthquake. iReport: Looking for loved ones in Haiti. Martino tried to sing a few verses. "There are so many things going through my head," he said. Overcome by emotion, he had to compose himself and start again. "People are saying Haiti is finished, but no, no, no, we will rise up," he sang softly. "We will strive. We will rebuild Haiti. We will stand united." "Don't be discouraged. There will be light at the end of the tunnel. My Haitian people. I will not let go. " Proceeds from downloads of the song are going to the nonprofit organizations Sow A Seed and MedShare. Music, Martino said, was so essential to Haitian life. But he didn't know when this rare silence would end; when he would be able to write lyrics, put them to melody. "We're all so traumatized," he said. Carnival, he said, was Haiti's musical showcase. "We've lost our biggest tradition. Carnival was part of us.""

Article number 66

"(CNN) -- She's the daughter of a Beatle, fashion designer of everyone from Madonna to the British Olympic team, recipient of a medal from Queen Elizabeth, and counts Kate Moss

among her friends. From all appearances, Stella McCartney's life has been a charmed one -- not that she necessarily sees it that way. Fame has followed the second child of Paul McCartney and American photographer wife Linda from the moment she was born. And it's been both a blessing and a hindrance. As McCartney was awarded this year's prestigious Women's Leadership Award from the Lincoln Center Corporate Fund, CNN spoke to the mother-of-four about sustainability in a notoriously "unaccountable" fashion industry. Parents' influence. CNN: You had very famous role models -- do you think that helped you, or do you think that was actually a hurdle? McCartney: It certainly opened a lot of doors and certainly closed some minds. So I think there was a balance. CNN: What do you think was the biggest inspiration you got from your mother, Linda McCartney? M: I learned a lot from her ethics. Both my mum and dad [former Beatle, Paul McCartney] are known to be vegetarians, world rights activists and environmentalists, and that definitely came into my place of work. Once you have children it adds another layer of responsibility to what you're doing. You have to be a role model to them so it makes you question your actions -- in a good way. Biggest Challenge. CNN: What is the biggest challenge you've encountered getting where you are now? M: Early on, when I wanted to go back to London and start my own fashion house, a very well thought-of executive in the industry said to me: "Name one female designer that's come from Great Britain that has had any kind of global success." I wanted to prove him wrong. Obviously, there have been great women from Britain in design, but actually there are fewer than I thought. So that was a bit of a hurdle for me. From day one I've never worked with leather or fur. I don't work with PVC, and I'm very conscious in the sourcing and manufacturing of fashion. That's a hurdle, that's a challenge, but it's a worthwhile one. CNN: Do you feel a lot of pressure from the industry because you chose to go down the "sustainable fashion" road? M: It's been difficult, it continues to be difficult, but I'm OK with difficult, it's what keeps me on my toes. I've had people say to me: "You'll never sell handbags, you don't work with leather and leather is luxury." To me it's the complete opposite, leather is everywhere, it's so cheap a material, it's so mass produced. Over 50 million animals a year are killed just for fashion. For me it doesn't have a luxury element to it. CNN: So you're comfortable flipping that notion on its head, that leather is luxury? M: Lots of things have been around for a long time, that doesn't mean they have to stick around forever. When you're in design it's your job is to change, to push, to modernize. The fashion industry, is not really as accountable for some reason. We're not expected to have to answer to the fact that it's not sustainable to kill that many animals for shoes and bags. And it's not necessary! Ninety percent of the people who come to my stores have no idea I don't work with leather. Being told no. CNN: You get very steely when you talk about people telling you you can't do something. M: Doesn't everyone, who likes being told they can't do something? Anyone can do anything they want, if they really want it. I'm not going to pretend I didn't come from a privileged starting point. I'll always admit that it was easier for me to question people telling me I couldn't do something because I had a pretty nice place to fall back on. So maybe that afforded me a little more spark and fight. I'm learning as I get older, that you don't have to try to fight everything from a man's place. I guess I got here for a reason, it wasn't because I was a woman. More: Bobbi Brown's billion-dollar idea."

Article number 18

"The "Three Amigos" -- or, as they're more formally known, the leaders of the United States, Canada and Mexico -- announced an agreement Wednesday to work on a plan to streamline trade and travel, including border controls among the countries. U.S. President Barack Obama, Canadian Prime Minister Stephen Harper and Mexican President Enrique Pena Nieto agreed to streamline border controls to facilitate the movement of people through the establishment of

a trusted traveler program. They also spent a great deal of time during the North American Leaders' Summit discussing efforts to broker a new trade agreement with Pacific nations, they said in a joint news conference after the summit in Toluca, Mexico. Obama has called for "fast-track" trade authority from Congress for him to pursue the so-called Trans-Pacific Partnership, a massive free-trade zone. But members of his own party, including House Minority Leader Nancy Pelosi and Senate Majority Leader Harry Reid, have voiced firm opposition to such authority. Republicans seized on Vice President Joe Biden's reported comments at last week's House Democratic retreat in Maryland. There, he was heard conceding the trade issue was quickly becoming a source of frustration within the party, especially among labor groups that are key to midterms next fall. "The jobs they seem to care about most are Democrats in Congress -- not families across the country eager to join the ranks of the employed," Senate Minority Leader Mitch McConnell said in a statement. Senior Obama administration officials played down reports about Biden's comments as coming from "second-hand accounts." But, they added, the White House remains firmly committed to its trade agenda. "It would not be in the interest of the United States to put this on the back burner," one official said. White House spokesman Jay Carney insists that differences among Democrats over trade issues date back several administrations. "The differing opinions on these matters are not new, and the fact that there are differing opinions within both parties is not new," Carney said Tuesday. Another area of friendly disagreement for the "Three Amigos" is over the Obama administration's handling of the Keystone XL pipeline. Canadian officials have grown impatient with the lengthy approval process in the United States for the contested project, which would transport oil from Alberta to the Gulf of Mexico. While a recent State Department environmental impact study appeared to brighten prospects for approval, senior administration officials indicated Harper is not likely to receive the news many in his nation want to hear during the summit. "I think what President Obama will do is explain to him where we are in the review of the Keystone pipeline, and indicate that we'll, of course, let our Canadian friends know when we've arrived at a decision," a senior administration official said. The Keystone project has also divided Democrats, namely environmentalists who see the pipeline as a symbolic battle in the larger fight over efforts to deal with climate change. While in Mexico, Obama faced another delicate balancing act over the issue of immigration reform. The plight of undocumented immigrants in the United States, notably the substantial number of migrants who crossed the border from Mexico, is a major political issue south of the Rio Grande. Latino special interest groups, a core Democratic Party constituency, have criticized the administration for its aggressive deportation policy as it struggles to find a long-term solution to the immigration issue. Senior administration officials told reporters last week that Obama remains committed to comprehensive immigration reform that includes a pathway to citizenship for undocumented workers. But that legislative priority has hit a roadblock in Congress, where Republican leaders have indicated there is little hope for a breakthrough this year before the midterms. "With respect to immigration, I think President Peña Nieto has a very good understanding, frankly, of the state of play in the United States," a senior administration official said."

Article number 71

"(CNN) -- There is nothing terribly surprising in Apple's refreshed line of tablets, but that's OK. We spent some time testing and touching the new iPad Mini and iPad Air after Tuesday's press conference. As promised by Apple executives, the new devices were lighter, thinner and seemingly faster -- just like many incremental product upgrades from the past. iPad Air. There were no major new features, such as the fingerprint scanner or camera upgrade that came with the iPhone 5S. The most unexpected news of the day was a new name for the $499 fifth-

generation iPad, which is now the iPad Air. The iPad line has been a bit wishy-washy with names. It began by counting each version, but then dropped the number and asked only to be know as "iPad," like Cher. The iPad 2 kept its number and a spot in stores, where its slightly lower price tag (now $399) might appeal to someone considering a cheaper Android or Windows tablet. Physically, the iPad Air is indeed lighter and thinner than its predecessor. It's 20% thinner than the third generation iPad, measuring in at a slight .29-inch. An ad shown at the press conference showed an iPad Air lying flat on a tablet, hiding discretely behind a No. 2 pencil. It has also dropped a bit of weight and the Wi-Fi version is now exactly 1 pound (the cellular version is 1.05 pounds). Unfortunately, I didn't have an older iPad on hand to do a weight comparison, and I'd picked up an iPad Mini first. Nothing makes a regular iPad feel hefty like holding an iPad Mini. You can't unfeel a Mini; it's just so delightfully wee. The iPad Air has a 64-bit A7 processor, the same chip recently introduced in the iPhone 5S. The A7 should benefit graphics heavy programs such as iMovie and iPhoto, and Phil Schiller claimed it would double the performance of the previous chip. It is difficult to judge speed increases during a few minutes in a crowded room without proper tests and an equally empty previous generation device to use for comparison. Fresh Apple products always feel zippier than previous generations. Part of that is the steadily improving processors inside, but its also the benefit of working on a new device that hasn't been gradually slowed down by pages of apps, hundreds of cat videos and the latest operating system upgrade. iPad Mini with Retina Display. The iPad Mini got its first upgrade since it was first launched one year ago. Physically, the device is a dead ringer for the original. Fire it up, and you'll see that it addresses one of the biggest complaints about the first version by upgrading the screen to a 2048-by-1536 display that packs in 326 pixels per inch, a big jump from the previous 163 pixels per inch. A number of Apple products already have the Retina Display: the iPhone 4S and up, the fourth generation iPod Touch, the 10-inch iPad and some MacBook Pros. (Apple says Retina Display is so good that the human eye can't spot out individual pixels on the device). Like the iPad Air, the iPad Mini with Retina has the A7 chip inside. Previously the Mini was a generation behind its larger iPad sibling in processor speed. But this upgrade and the new display put the devices on the same level, so picking a smaller screen no longer means opting for the inferior product. The smaller tablet size is a competitive market, and the iPad Mini with Retina Display is finally Apple's premium offering for the 7.9-inch size. The $399 tablet is now in a better position to take on products such as Google's Nexus 7, Samsung's Galaxy tablets and Amazon's Kindle HDX. Though other products can usually be had for much less, any hands-on time with Apple products remind you that one of their greatest selling points is solid, high-quality construction. The missing upgrade. Shaving off excess bulk and swapping out faster processors is de rigueur for any product update these days. Other companies might throw in price drops, but Apple is comfortable in its spot at the high-end of the consumer market. It won't start slashing devices for its flagship devices. Instead, the company keeps older models available, though even those are priced somewhat high. Unfortunately, there is one key feature where all the tablet makers seem to have stalled: battery life. Faster processors gobble up more battery power, so maintaining 10 hours of battery life from an old tablet to a new tablet probably requires battery improvements. Both new iPads have the same M7 motion co-processor that handles sensors, such as the accelerometer and gyroscope, saving battery power by not using the main processor. Other than an hour here or there, mobile devices makers seem stuck at the half-day mark, a frustrating limitation for people who have adapted by carting around chargers and scrambling to re-up anytime they spot an empty outlet at a cafe, airport or friend's house. For now, Apple is comfortable enough with the competition to just iterate on its well-designed, popular products, making the usual round of improvements and throwing in the occasional flashy feature to grab attention, such as fingerprints on the iPhone 5S. It may not be as much fun for tech fans who were used to splashy new products and well-kept secrets, but it's probably a solid

business decision. According to Forrester analyst Sarah Rotman Epps, replacement tablet sales are growing faster than new tablet sales worldwide. Manufacturers need to compete to keep the customers they already have who are considering upgrading from an old model. All the tablet makers are competing with what are essentially the same minor tweaks. The competition will stay close until the first company breaks out of the usual faster, thinner, faster upgrade cycle and solves the battery problem."