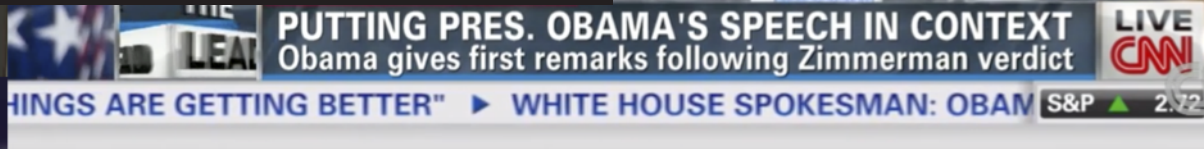


# Multimodal information extraction from videos

Automatic creation of highlight clips from political speeches

Ombretta Strafforello



Technische Universiteit Delft





# MULTIMODAL INFORMATION EXTRACTION FROM VIDEOS

AUTOMATIC CREATION OF HIGHLIGHT CLIPS FROM POLITICAL  
SPEECHES

by

**Ombretta Strafforello**

in partial fulfillment of the requirements for the degree of

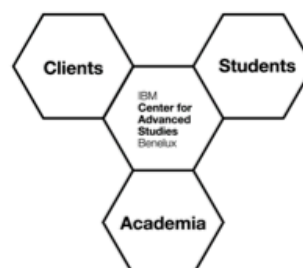
**Master of Science**  
in Computer Science

at the Delft University of Technology,  
to be defended publicly on Thursday October 3, 2019 at 13:00.

Supervisor: Dr. N. Yorke-Smith  
Thesis committee: Drs. O. Inel, TU Delft  
Prof. dr. M. Loog, TU Delft  
Drs. B. Timmermans, IBM CAS Benelux

*This thesis is confidential and cannot be made public until October 3, 2020.*

An electronic version of this thesis is available at <http://repository.tudelft.nl/>.





# ABSTRACT

With the huge amount of data that is collected every day and shared on the internet, many recent studies have focused on methods to make multimedia browsing simple and efficient, investigating techniques for automatic multimedia analysis. This work specifically delves into the case of information extraction from videos, which is still an open challenge due to the combination of their semantic complexity and dynamic nature. The majority of the existing solutions are tailored for specific video categories and result in the creation of key frames time-lapses, video summaries, video overviews or highlight clips. In particular, this thesis project focuses on the case of highlights extraction from videos where one person speaks facing the camera. Automating the process of analysis of this specific kind of videos is important in the industrial context because it can be harnessed for several interesting applications, such as the automatic video summarisation of interviews or the automatic creation of personal video curricula vitae.

In this setting, the research objective is to investigate how Machine Learning can be deployed for the task of information extraction. Several options are possible, as from the target videos multiple types of features can be extracted, such as textual features from the speech transcription; visual features from the facial expressions, head pose, eye gaze and hand gestures; audio features from the variations in the tone of the voice. The exploitation of multimodal features enhances the capacity of Machine Learning algorithms. In fact, as proven in former research [1–10], the integration of multiple channels of information — textual, audio, visual — makes it possible to derive a more precise and greater amount of knowledge, just like humans exploit their multiple senses, in addition to experience, to make classifications or predictions. In this work, two approaches for multimodal information extraction from videos are investigated. The first approach is based on simple multimodal feature vectors concatenation, inspired by [11], while the second approach exploits a recent deep architecture, the Memory Fusion Network [12], to model both individual and combined temporal dynamics. To test the effectiveness of multimodal learning in the context of information extraction from videos, the two techniques are compared against a unimodal, content-based method, that relies on the summarisation of the video transcripts.

In order to train the multimodal approaches in a supervised fashion, a novel dataset based on videos of political speeches of well-known American politicians, the *Political Speeches Dataset*, was collected. The dataset is provided with binary saliency labels, that allow to identify the ground truth salient video segments. Four types of highlight clips are generated for each speech and evaluated through crowdsourcing. The results show that the quality of automatically created highlight clips is comparable to the ground truth, in terms of informativeness and ability to generate interest. Moreover, they also confirm that highlight clips generated with multimodal learning are more informative than the baseline.



# PREFACE

One year ago, Dr. Zoltán Szlávik and Drs. Oana Inel offered me the possibility to become a research intern at IBM Center for Advanced Studies. There, I had the pleasure to meet fantastic fellow graduating students, and be part of a stimulating environment, where each of use was involved in a different research concerning emerging fields in Artificial Intelligence.

In my case, I accepted the challenge of automatic multimedia analysis, objective of the *Avengers Project* and topic of my M.Sc. thesis. Automatic multimedia analysis remains an open problem: even though former research successfully produced algorithms that extract information from specific types of multimedia content, for example, in the case of text mining or image recognition, a unified system that process and extracts relevant information from a generic kind of input is still to be invented. With regard to this, the Avengers Project team designed a pipeline architecture that makes it possible to organise the multimedia analysis in a modular way and process each input modality separately. In this context, I delved my work into a particular input type: videos of people that speak in front of a camera.

Previous research on automatic video analysis and summarisation revealed the complexity of this problem, and the limitations of current Machine Learning models to bridge the semantic gap. For this reason, at first I felt discouraged, for I thought making a scientific contribution in this field was a daunting challenge. However, during this journey I found the support of many people, that inspired me with their suggestions, and made the working time more pleasant with their company.

Firstly, I am deeply thankful to my thesis supervisors, for always giving me the possibility to conduct the research in my own fashion, supporting my choices and believing in my ideas. During the thesis, I could freely put my knowledge into practice, organise the work in an independent manner, making my own decisions and expressing my creativity.

Dr. Neil York-Smith, thank you for your constant support, your availability and your precious advice.

To my supervisors from IBM CAS, Drs. Oana Inel, Drs. Benjamin Timmermans and Dr. Zoltán Szlávik, thank you for offering me the opportunity to experience scientific research both from an academic and industrial perspective. More importantly, thanks to you I was introduced to the topic of information extraction from videos with Machine Learning, that deeply fascinated me and shaped my future career.

Thanks to the fellow interns from IBM CAS, for the amusing time between the working hours and your support.

I am very grateful for my fellow classmates and friends from TU Delft, who shared with me long hours in the library, mutual encouragement and support. Thanks for being amazing friends.

Thanks to my longtime friend Federica, who kept me good company in the last days of thesis writing.

Most importantly, I would like to thank Jonathan for always being by my side during this year, both in happy and difficult times.

Lastly, I would like to thank my parents, who are going to fly to Netherlands to be present for my thesis defense.

*Ombretta Strafforello  
Delft, September 2019*





# CONTENTS

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation and objectives . . . . .	2
1.2	Research questions . . . . .	3
1.3	Contributions . . . . .	3
1.4	The Avengers Project . . . . .	4
1.5	Thesis outline . . . . .	5
<b>2</b>	<b>Literature review</b>	<b>7</b>
2.1	Introduction . . . . .	7
2.1.1	Harnessing A.I. for Augmenting Creativity . . . . .	7
2.1.2	Tropes . . . . .	8
2.2	Video summarisation . . . . .	8
2.2.1	Methodology to search for papers . . . . .	9
2.2.2	First approach to automatic video summarisation . . . . .	9
2.2.3	Overview of recent supervised and unsupervised methods . . . . .	10
2.2.4	State-of-the-art . . . . .	11
2.2.5	Discussion . . . . .	13
2.3	Video segmentation . . . . .	13
2.3.1	Methodology to search for papers . . . . .	13
2.3.2	Kernel temporal segmentation . . . . .	13
2.3.3	Shot boundary detection problem . . . . .	14
2.3.4	Audio segmentation . . . . .	14
2.3.5	Online change point detection . . . . .	14
2.3.6	Discussion . . . . .	15
2.4	Multimodal learning and its applications . . . . .	15
2.4.1	Methodology to search for papers . . . . .	15
2.4.2	Examples of multimodal Machine Learning . . . . .	15
2.4.3	Multimodal learning in social signal processing . . . . .	16
2.4.4	Memory Fusion Network . . . . .	17
2.4.5	Discussion . . . . .	18
2.5	Multimodal datasets . . . . .	18
2.5.1	Methodology to search for papers . . . . .	18
2.5.2	Multimodal datasets . . . . .	18
2.5.3	The MOSI and MOSEI datasets . . . . .	19
2.5.4	Discussion . . . . .	20
2.6	Crowdsourcing . . . . .	20
2.6.1	Methodology to search for papers . . . . .	20
2.6.2	The use of crowdsourcing in support of Machine Learning . . . . .	20
2.6.3	Discussion . . . . .	21
2.7	Conclusion . . . . .	22
<b>3</b>	<b>Methodology and models for the highlights extraction</b>	<b>23</b>
3.1	Introduction . . . . .	23
3.2	Dataset . . . . .	24
3.3	Machine Learning methods for highlights extraction . . . . .	25
3.3.1	Baseline: Text summarization . . . . .	25
3.3.2	Multimodal features concatenation . . . . .	26
3.3.3	Multimodal features integration with MFNs . . . . .	26

3.4	Evaluation . . . . .	26
3.4.1	Automatic evaluation . . . . .	27
3.4.2	Evaluation with crowdsourcing . . . . .	27
<b>4</b>	<b>Novel dataset for highlights extraction from political speeches</b>	<b>29</b>
4.1	Motivation . . . . .	29
4.2	Creation of the dataset . . . . .	30
4.2.1	Labelling process . . . . .	31
4.2.2	Multimodal features extraction . . . . .	32
4.3	Content of the dataset. . . . .	33
<b>5</b>	<b>Experiments</b>	<b>35</b>
5.1	Introduction . . . . .	35
5.2	Preprocessing of the data . . . . .	35
5.2.1	Data dimensionality reduction. . . . .	36
5.3	Baseline method . . . . .	39
5.4	Machine Learning models with multimodal features concatenation . . . . .	40
5.4.1	Mini-Batch K-Means clustering . . . . .	40
5.4.2	Density-Based Spatial Clustering of Applications with Noise. . . . .	42
5.4.3	Random Forest. . . . .	45
5.5	Machine Learning models with multimodal features integration . . . . .	46
5.5.1	Input data . . . . .	47
5.5.2	Modifications in the code . . . . .	48
5.5.3	Hyperparameters . . . . .	49
5.5.4	First results . . . . .	49
5.5.5	Data augmentation . . . . .	49
<b>6</b>	<b>Crowdsourcing methodology</b>	<b>53</b>
6.1	Introduction . . . . .	53
6.2	Evaluation with crowdsourcing . . . . .	53
6.2.1	Crowd . . . . .	54
6.2.2	Initiator . . . . .	55
6.2.3	Process. . . . .	55
6.3	Pilot surveys . . . . .	56
6.4	Estimation of the number of crowdworkers required . . . . .	56
6.4.1	Individual assessment . . . . .	56
6.4.2	Pairwise comparison. . . . .	58
6.5	Estimation of the budget . . . . .	60
6.5.1	Individual assessment . . . . .	60
6.5.2	Pairwise comparison. . . . .	60
6.6	Preliminary results . . . . .	61
<b>7</b>	<b>Results and discussion</b>	<b>63</b>
7.1	Introduction . . . . .	63
7.2	Automatic evaluation . . . . .	63
7.3	Evaluation with crowdsourcing . . . . .	65
7.3.1	Individual assessment . . . . .	65
7.3.2	Pairwise comparison. . . . .	71
7.4	Discussion . . . . .	75
<b>8</b>	<b>Conclusion and future work</b>	<b>77</b>
8.1	Considerations and conclusions . . . . .	77
8.2	Answers to the research questions . . . . .	78
8.3	Recommendations and suggestions for future work. . . . .	80
8.3.1	Experiment with more structured video types . . . . .	80
8.3.2	Enlarge the Political Speeches Dataset . . . . .	80
8.3.3	Extend the MFN . . . . .	80
8.3.4	Expand the crowdsourcing study. . . . .	80

---

<b>Bibliography</b>	<b>81</b>
<b>A Political Speeches Dataset</b>	<b>89</b>
A.1 Full list of political speeches from the dataset . . . . .	89
A.2 Missing words using "Common Crawl" for the pre-trained word vectors . . . . .	100
<b>B Crowdsourcing study</b>	<b>103</b>
B.1 Highlights presented in the evaluations . . . . .	103
B.2 Figure Eight surveys . . . . .	104



# 1

## INTRODUCTION

In the digital era, a huge amount of data is produced every day, both by individuals, that take pictures, generate videos or record vocal messages, both in the entertainment business, as in the case of media broadcasting or the music industry. These multimedia data are often shared on the internet, especially on social media or on famous media-sharing platforms like YouTube. Given this context, many recent studies have focused on methods to make multimedia browsing simple and efficient [13–17]. The solutions consist in information retrieval systems that make it possible to easily extract relevant content from unstructured text [18–20] or quickly identify images that match specific queries [21–23].

In the case of videos the problem is harder, due to the intrinsic complexity of this type of data, given by the combination of its convoluted semantics and dynamic nature. What is more, videos are characterised by the presence of multiple channels of information: visual information from the frames content, audio information from the possible presence of soundtracks and other auditory patterns, and textual information from the speech transcripts. An additional difficulty is brought by the fact that analysing these features without taking into consideration the temporal dimension is limiting, as information about the cause-effect relations would be lost. Because of all these aspects, existing previous research is tailored for the analysis of specific video categories, from which it is possible to forecast all types of actions and dynamics that might occur. For instance, there exist several works concerning sports highlights extraction [6, 24–28]. The case of sports videos offers a lot of advantages that make the automated analysis easier: they follow a predetermined scheme, with a fixed set of characters with precise roles in the game, a series of possible actions constrained by the rules and a recurring vocabulary of domain-specific terms and expressions, used by the commentators and spectators.

Most of the solutions created to convey the important information extracted from videos consist of key frames collections [29, 30], video summaries [31, 32], video overviews [33] or highlight clips [6, 27]. This thesis project focuses on the particular case of highlights extraction from videos where one person speaks facing the camera. Among other reasons, automating the process of analysis of this specific kind of videos is important in the industrial context because it can be harnessed for several interesting applications, such as the automatic video summarisation of interviews or the automatic creation of personal video curricula vitae. This kind of videos have been extensively investigated in the field of social signal processing, for the purpose of recognising the emotions and the sentiment expressed by the subjects [12, 34–37]. However, the research purpose of this thesis differs from the former example, as it represents the first attempt to identify salient content and use it so to create highlight clips.

In this setting, the research objective is to investigate how Machine Learning can be deployed for the task of information extraction in the specific case of videos. In fact, since multiple types of features can be identified from the target videos, several models can be designed, each based on a specific subset of these. The types of features involved may be: textual features from the speech transcript; visual features from the facial expressions, head pose, eye gaze and hand gestures; audio features from the variations in the tone of the voice. The research is conducted under the hypothesis that not only Machine Learning can be used successfully for the task of automatic highlights extractions, but also that the exploitation of multimodal features enhances its capacity. In fact, as proven in former research [1–10], the integration of multiple channels of information — textual, audio, visual — makes it possible to derive a more precise and greater amount of knowledge, just like humans exploit their multiple senses, in addition to experience, to make classifications or predictions.

The results of this thesis project confirm that highlight clips generated through multimodal models are perceived by human judgment as more informative and more able to generate interest than highlights created by professional filmmakers and by unimodal Machine Learning based on speech transcripts only.

### 1.1. MOTIVATION AND OBJECTIVES

The central objective of the thesis is the **discovery of an effective Machine Learning method for the extraction of highlights from videos where one person speaks to the camera**. This starts from an automatised video analysis that can be conducted on more levels. In fact, it is possible to perform a purely content-based analysis and apply text summarisation techniques in order to detect and understand relevant information. However, from these videos it is possible to identify additional features, namely the ones regarding the facial expressions of the speakers and their tone of the voice.

Multimodal is a key word in this thesis. The intuition behind the decision of using multimodal Machine Learning is matured from the idea that a comprehensive feature representation leads to better accuracy in the results of Machine Learning. To this regard, the main hypothesis of this research is formulated.

**Hp:** *Multimodal features can be used to train Machine Learning models that give better results in information extraction from videos, compared against unimodal methods.*

However, this needs to be verified, as an approach purely based on the content, that is the videos transcription, might be still more effective in terms of quality of the information extracted. In fact, the inclusion of more information channels might only disturb the classification with the introduction of noise. In addition, several possibilities for the aggregation and jointly usage of multimodal features exist. It is one of the objectives of this thesis to investigate what are the best methods of integration of features from different modalities that make it possible to achieve outstanding results, possibly better than those achieved using a single feature modality.

The research on the analysis of videos where one person speaks to the camera led to discovery of the MOSI and MOSEI datasets [37, 38], a collection of movie reviews from YouTube with sentiment intensity and emotion annotation, in addition to already computed text, audio and video descriptors. Initially, the project was built around this dataset, with the ambition of utilizing both the content information from the video transcripts and multimodal sentiment analysis to extract salient parts of the reviews in an unsupervised fashion. For example, the key fragments of the video reviews where the subject expresses either appreciation or dislike can be identified by changes of the tone of the voice and the use of certain facial expressions, like a smile or, on the contrary, a grimace. However, the problem of the evaluation of the results of this unsupervised approach represented an obstacle for its realisation. In fact, because of the lack of annotations concerning the relevance of each video segment, crowdsourcing was left as the only possibility for the evaluation. However, designing an effective crowdsourcing task under the available budget and time constraints was something unfeasible.

This led to the redesign of the thesis case study, developed around the creation of a novel multimodal dataset. Sticking with the definition of the target video category, in which one person speaks facing the camera, the research for useful data led to a sharper focus on videos of political speeches, that are usually recorder and shared on the Internet. Since in most case these videos are centered around one main politician that is speaking to the public, in front of one or more cameras, this type of videos was chosen as a good candidate for a possible dataset usable for the thesis objective. To this end, a dataset composed of videos from 99 political speeches was created and named "Political Speeches Dataset". This dataset is mentioned throughout the thesis and is described in detail in Chapter 4. What makes the "Political Speeches Dataset" unique is the fact that it contains 15624 independent video segments labelled according to their relevance in the context of the original speech. The term "relevance" or "saliency" are used in this case to indicate video segments from their respective speech that are important because they are *likely to be included in a highlight clip created by a news channel*. The *saliency labels* can be deployed for the training of supervised Machine Learning and Deep Learning algorithms, used for the extraction identification of relevant moments in a video. The "Political Speeches Dataset" represents the very first attempt of creation of a labelled dataset for the task of automatic video understanding of videos where one person speaks to the camera and it constitute the first contribution of this thesis project.

In order to investigate different ways of deployment of multimodal features for the task of automatic information extraction from videos where one person speaks facing the camera, two distinct approaches for highlight clips creation were developed, under the influence of the papers "Harnessing A.I. for augmenting

creativity: Application to movie trailer creation", a former IBM project [11], and "Memory fusion network for multi-view sequential learning" [12]. The first paper suggested to tackle the problem of scene classification from a multimodal perspective, including multiple channels of information, such as those represented by sounds and images. In this work, the multimodal features are simply concatenated to form a single feature matrix, that is fed to a classification model. This simple techniques for the multimodal feature integration might lead to suboptimal results. Therefore, a second approach was attempted, namely the Memory Fusion Network (MFN) proposed in [12]. In the MFN, the multimodal features are processed individually and then correlations among the different modalities are detected and used as additional features for the classification. In theory, this allows the complete exploitation of the multimodal setup, thus extracting all the information conveyed by a single channel and by the combination of different channels together. Understanding which technique is more suitable for the task of video classification with multimodal Machine Learning represents one of the thesis objectives.

Finally, the evaluation of the results of information extraction from videos represents a true challenge in this research. As already mentioned, it is hard to assess which moments are important in a video where one person is speaking without the intervention of a human. If the "saliency labels" introduced in the Political Speeches Dataset allow the training of supervised Machine Learning algorithm, the highlight clips resulting from these methods need to be evaluated by human individuals. It is the ultimate objective of this research to implement a thorough human evaluation of automatically generated highlight clips through a crowdsourcing process, designed in a way to avoid human bias and gather significant answers.

## 1.2. RESEARCH QUESTIONS

Having defined the general research purpose, the thesis is going to be structure in order to address three main research questions, with the relative research subquestions.

**RQ1** *Do the salient moments of a video where one person speaks facing the camera share common properties, i.e. the recurrence of particular images, actions, sounds or verbal expressions, that can be identified, classified and used to categorize the scenes of the video?*

- *If they exist and are identifiable, can the salient scenes contained in a video where one person speaks facing the camera be used to represent the relevant information contained in the video? For example, can this be done through the creation of a short highlight clip that shows the most significant moments?*

**RQ2** *If the research question RQ1 is verified as true, is it possible to automatize the process of relevant information extraction from videos where one person speaks facing the camera using Machine Learning and Deep Learning techniques?*

- *When using Machine Learning for video analysis with the purpose of information extraction, do the combination of textual, audio and visual features outperform purely content-based methods?*
- *If so, what is the most effective method for the integration of multimodal features?*

**RQ3** *Is crowdsourcing an effective method for the evaluation of the results of automatic information extraction from videos?*

- *If crowdsourcing can be used for the evaluation of the results, how should the tasks included in the crowdsourcing process be designed?*

## 1.3. CONTRIBUTIONS

The main contributions of this thesis can be summarised as follows.

- A novel dataset for the task of relevant information detection from video recorded political speeches, namely the *Political Speeches Dataset*, was collected. The dataset involves 99 political speeches from six well known American politicians, divided into 15624 video segments. Each video segment is provided with a binary "saliency label", that reflects the probability of the segment to be chosen for a highlight clip. The dataset offers a multimodal features representations, namely a collection of text, audio and video features that were extracted from each video segment.

- The methodology for movie scenes classification first employed by Smith, John R., et al. [11], is here replicated and extended to the new applications of scene saliency prediction, demonstrating its efficacy with a new set of multimodal features.
- The novel architecture introduced in [12] for the purpose of multimodal sentiment analysis, namely the *Memory Fusion Network*, is replicated and adapted in this thesis for the binary classification of "salient" and "non salient" video segments.
- The thesis includes a complete benchmarking which involves both supervised and unsupervised, as well as unimodal and multimodal, Machine Learning methods for the task of video segments classification
- The experimentation with different methods for the extraction of highlights from videos led to the creation of four types of highlight clips for each speech belonging to the Political Speeches Dataset, which are included in the dataset and can be used for future research
- Due to the lack of tool or a metric for the automatic evaluation of the results of relevant information extraction, a thorough crowdsourcing user study, composed of two crowdsourcing jobs, was designed in order to achieve a complete and unbiased human evaluation of the extracted highlight clips.
- Finally, the thesis includes an analysis of the limits of the research on information extraction from videos and recommendations for future work.

#### 1.4. THE AVENGERS PROJECT

This Master's thesis is fruit of a collaboration between TU Delft and IBM Center for Advanced Studies Benelux<sup>1</sup>. Every year IBM CAS establishes collaboration with Dutch universities in order to conduct research on the most innovative topics in information technology. The present study is part of a project that has the objective of building a tool capable of extracting the most relevant information from various multimodal input, consisting of text files, audio files, pictures or videos. The extracted content can be further elaborated and aggregated in order to produce a final video of short duration that gives an overview of the most salient information contained in the multimodal input, like a highlight clip. Because of its nature, that is to some extent based on a creative production, the project was named the "Avengers Project".

The team working on the Avengers Project designed a pipeline architecture that makes it possible to visualise and organise the target system in a modular way. As can be seen in Figure 1.1, according to the design, each input mode is processed by a different block in the pipeline. The results of each block are then recombined together to produce the output, either automatically or by the user. Four types of input modes were distinguished:

- **Text:** a generic text file, e.g. a review, a plot, a curriculum vitae;
- **Audio:** an audio file, e.g. a soundtrack or the record of a speech;
- **Video I:** film of one single subject speaking facing the camera, possibly in a close-up, for example as in a YouTube review;
- **Video II:** any type of video that is not already included in video type I.

<sup>1</sup><https://www.research.ibm.com/university/cas/benelux/>



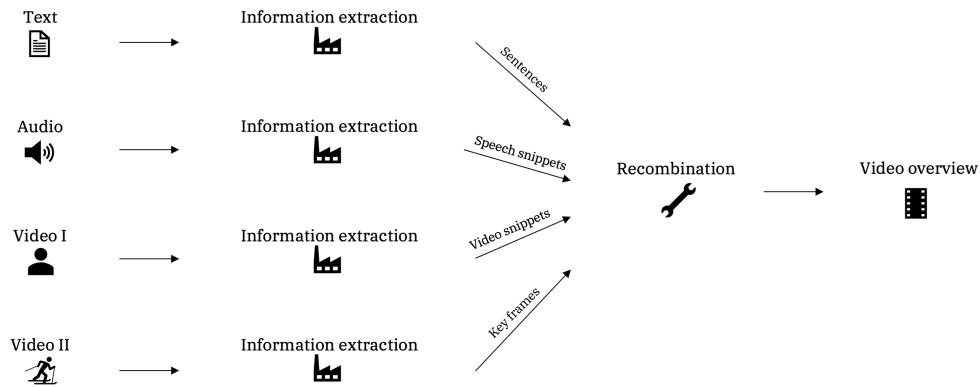


Figure 1.1: Pipeline design for the Avengers Project

As already mentioned, the target of this thesis are the videos of type I, namely close-ups where a single person is filmed while they are speaking to the camera, such as in a monologue, a YouTube review or an interview. The investigation conducted on this specific kind of videos started from an extensive research of the literature, mainly concerning video summarisation and multimodal learning, which led to the design of a guideline framework, visible in Figure 1.2, where the process for the creation of the final video overview is decomposed into four basic steps.

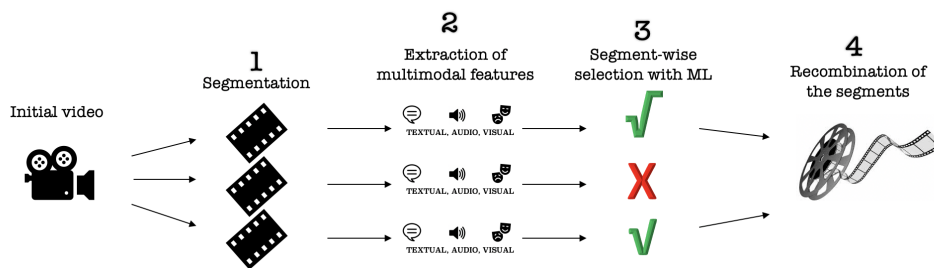


Figure 1.2: Basic structure of the video overview creation process

The first step consists in splitting the videos into consecutive segments. The temporal segmentation is done according to the pauses in the speech. The video analysis is then conducted at a segment level: from each segment, multimodal features are extracted, namely descriptors for the audio, textual and visual content. The multimodal features are then fed to a classification model that predicts the relevance of the video segment in context of the full video. Finally, the video segment classified as relevant are concatenated to produce the final highlights video. Mindful of the objective of the Avengers Project, the general framework here introduced is shared by all the methods that were experimented in this research.

## 1.5. THESIS OUTLINE

The investigation of the literature review that inspired the objectives of the thesis and the definition of the research questions is presented in Chapter 2. It is followed by the description of the methodology, which provides the content of Chapter 3. Chapter 4 introduces the novel "Political Speeches Dataset" and explains how it was created and how it is structured. Consequently, Chapter 5 presents the experiment that were conducted in order to discover the best performing Machine Learning models. The results obtained by these methods are evaluated through a crowdsourcing process, which is defined in Chapter 6 from the analysis obtained from experimenting with a pilot crowdsourcing round. The final results of this research are presented in Chapter 7, where they are analysed from a low level and a high level perspective. Finally, Chapter 8 closes this dissertation presenting the final general conclusions, as well as some recommendation for future research.



# 2

## LITERATURE REVIEW

### 2.1. INTRODUCTION

This chapter presents the literature review that inspired the definition of the objectives, the creation of the methodology and the choice of the techniques utilised in this research. First, the paper that influenced IBM Center for Advanced Studies to continue the investigation of automatic multimedia analysis through Machine Learning is presented (§2.1.2). In §2.2, there follows an introduction to the researches on the topic of automatic video summarisation, which is closely linked to the problem of highlights extraction from videos, conducted over recent years. After that, §2.3 contains a brief introduction of methods usable for video segmentation. The literature review continues with an investigation of the most recent researches on multimodal Machine Learning, a domain that results fundamental to the present study; precisely, it offers relevant insights concerning the manners in which combinations of textual, audio and visual features that can be extracted from video can be used for classification problems (§2.4). Particular attention is drawn to the challenge of automatic sentiment analysis and emotion recognition, which led to a deep investigation of Machine Learning methods for video analysis and are, therefore, worth a consideration related to the problem of information extraction from videos. Related to the antecedent section, §2.5 presents a list of the existing multimodal datasets. Finally, the literature review ends with a section about crowdsourcing, which is an efficient strategy both for the collection of labelled data and for the evaluation of the research results.

#### 2.1.1. HARNESSING A.I. FOR AUGMENTING CREATIVITY

The paper that inspired this research on information extraction from videos is titled "Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation" and was written by a team of IBMers, including John R. Smith, Dhiraj Joshi and Jozef Cota, from the T. J. Watson Research Center in Yorktown Heights (NY, USA), in collaboration with EURECOM (Sophia Antipolis, France) and the National Taiwan University [11]. In their work, the authors prove the effectiveness of multimodal features in the task of movie trailers creation. In particular, they deal with the classification problem of predicting the scenes of horror movies that are likely to appear in their trailers, which are otherwise typically designed by professional filmmakers in a process that requires extended time resources and expertise.

The dataset that was used, includes 100 horror movie trailers, where scenes were segmented in order to produce audio and visual snippets. Since horror movie directors play with audio tracks, lights and colours to arouse certain emotions in the spectators, the multimodal descriptors that are extracted from the videos may be regarded as representative of both the content and the sentiment that belong to the scenes. From the audio snippets, an emotional vector representation is extracted with OpenEAR [39], an extension of the software OpenSMILE for audio analysis [40], while the visual snippets are represented by one key-frame descriptors. Visual descriptors include sentiment scores, computed with Sentibank [41] from the image content, and scene oriented visual attributes that describe the context globally, extracted with the model Places 205 CNN for scene recognition [42]. Visual and audio features are then aligned and aggregated by averaging the descriptors of the key-frames that fall within the same audio segment. The authors make use of this dataset to design a statistical framework that can be used to identify the recurrent properties of the scenes in horror movie trailers. To this end, Principal Component Analysis (PCA) is applied to reduce the dimensionality of the feature vectors, resulting in three main components that are sufficient to capture the essential character-

istics of the trailers scenes. The authors claim that these three dimension approximately correspond to the attributes "scary", "tender", and "suspenseful".

The model is tested on the movie Morgan, by 20<sup>th</sup> Century Fox<sup>1</sup>. Two trailers were produced for this movie: the original one by 20<sup>th</sup> Century Fox and one by a professional filmmaker who used the Augmented Intelligence system for scenes selection<sup>2</sup>. The analysis of the results showed that the two trailers had several scenes in common, indicating a human-like performance of the automatized system. Moreover, it displayed an ability of identifying the scenes that should appear in the common structure of a horror movie trailer, without revealing too much of the movie content. This entails a substantial time and costs saving. In addition, the trailers were evaluated through two separate user studies where the participants, fifty for each survey, were asked about their interest in horror movies and appreciation of the trailer, and whether they thought that the trailer was made by an AI system. The rating distributions of the two trailers both resulted in a bell shape and the Turing test showed that the users were not able to differentiate between the AI-made trailer and the 20<sup>th</sup> Century Fox trailer. The efficacy of this project proves the effectiveness of multimodal learning in the field of video understanding and shows how Artificial Intelligence can be used as a support for human creativity.

### 2.1.2. TROPES

Essentially, the ultimate goal of the authors is to become able to recognize and distinguish "tropes", i.e. storytelling devices that are frequent in film and TV productions. Tropes consist in a combination of meaningful objects, scenes, sounds and colors that suggest the spectators what is going to happen in the next scenes. For example, as reported in the paper, in sports videos two common tropes for describing the salient contexts are the "buzzer beater" and the "Hail Mary pass". The first refers to a winning shot in the last seconds of a basketball game, right before the buzzer beater, and the second to a long forward pass in American football, performed due to desperation of the player given its small chances of success. Another example is the recurrence of the trope "Oh, Crap!" in the film "Interstellar", used to announce that something bad is going to happen, as reported on the website TVTropes<sup>3</sup>. This moment is characterized by the use of stronger language, of facial expressions indicating realization/panic by the characters, and accompanied by exclamations like "You Have Got to Be Kidding Me!", "I Want My Mommy!" or "This Cannot Be!", as well as the choice of the first few notes of Fryderyk Chopin's Funeral March as background music.

As the researches described in the paper "Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation" proves that it is possible to push the boundary of multimodal learning for the investigation of movie scenes properties, this thesis project aims at investigating the power of multimodal feature in the task of political speech understanding and highlights extraction. This is motivated by the fact that, as it is possible to identify and distinguish tropes in TV productions, it might be possible to identify common underlying structures in political speeches, as well for recurrent oral techniques used by politicians to enhance their credibility or for drawing the attention when they are about to say something important. The identification of these can be exploited to extract highlights from videos of political speeches.

More information about multimodal learning and information extraction from videos will follow in the next sections, especially §2.2 and §2.4.

## 2.2. VIDEO SUMMARISATION

Even though this thesis tackles the problem of information extraction from videos by focusing on the case study of highlights extraction, it is necessary to introduce the topic of video summarisation, to which the former is strongly linked. The two problems differ in the sense that video summarisation has a stricter definition. A good quality summary is supposed to describe, to a certain extent, all the relevant moments happened in the video. A "highlights clip" contains a collection of salient moments that are, for some reason, interesting to the viewer. A highlight clip can still be considered of good quality even though it does not cover all the important parts from the original video, as long as it conveys some extent of information and engages the viewer. Nevertheless, the techniques adopted for video summarisation can be applied also for highlights extraction, although the criteria of evaluation of the latter can be less rigid.

<sup>1</sup><https://www.foxmovies.com/movies/morgan>

<sup>2</sup>The trailer created with IBM Watson is available at <https://www.youtube.com/watch?v=gJEzuYynaiw>

<sup>3</sup><https://tvtropes.org/pmwiki/pmwiki.php/Main/OhCrap>

### 2.2.1. METHODOLOGY TO SEARCH FOR PAPERS

The topic of video summarisation was originally inspired by the work on movie trailer creation carried out by Smith, John R., et al. at the IBM T. J. Watson Research Center [11]. The research for related work on automatic summarisation was conducted on Google Scholar [43], searching for papers matching the keywords "automatic video summarisation". Successively, the research was narrowed down using the words "video skimming", as this category is the one chosen for the output of the summariser developed in this thesis. Particular focus was put on systems involving neural networks, as cutting-edge deep learning models have shown promising results, even in the task of automatic summarisation [27, 31, 44–46]. Among these, recurrent neural networks were preferred, for their ability to model long-term dependencies. To search for related papers, keywords like "video summarisation recurrent neural networks" or "video summarisation LSTM" were utilized. In addition, the references of each paper were consulted, as well as the papers that referenced the papers in question. Finally, in order to find the cutting edge techniques, the keywords "video summarisation state-of-the-art" were typed, filtering out results published before 2018. All the citers of the remaining were also considered, in order to track down all the latest researches in this domain.

### 2.2.2. FIRST APPROACH TO AUTOMATIC VIDEO SUMMARISATION

Given an input video, video summarisation refers to the process of identifying the relevant information contained in the video, extracting it and recombining into a shorter video that conveys the original message. This task is not trivial. In fact, even though there are some straightforward properties that should describe every video summary, such as continuity (a smooth flow with no interruptions), priority (salient scenes should be prioritized) and repetition (the same scene should appear just once) [47], some further questions might be answered differently — such as how much shorter should a summary be and what is the important content that it should contain — depending on the particular source video or user.

The challenge of automatic video summarisation has been tackled since the early 2000s. From first attempts of video summarisers like VidSum [48], nowadays the advance of (Computer Vision and) Deep Learning has brought sophisticated solutions, that produce state-of-the-art convincing trailers and summaries [11, 46]. Such new techniques allow to automate the process of content identification and summary assembling by reducing the need of fixed schemes and structures and human intervention. On the contrary, a simpler tool like VidSum [48], which was used by the authors to summarise Forum presentations, relies on predetermined presentation structures (PSs) and summary design patterns (SDPs). More precisely, VidSum is formed by five crucial steps. Initially, the low-level features of the video, such as laughter, slide changes, cuts, speaker changes, etc., are detected. These are used to determine the PS, namely the layout that determine the sequence of events in the video. Knowing the PS, the best SDP, that is the template that describes the way video, audio and text are combined to produce the output, is selected, and elements from the original video are matched to the SDP slots. Once this is done, the video summary just needs some refinements, such as the addition of text and visual adjustment where necessary. This method produces high-quality results when dealing with static video settings, as in the case of Forum presentations. However, when the source video is scarcely structured VidSum fails to identify salient scenes.

In VidSum, the structure of the video summaries depends on the chosen SDP. In general, there are two types of video summaries: video skimming and key frame-based video representation. Video skimming refers to the technique of creating short synopses of a video segment by concatenating fragments of significant video shots with the respective audio, preserving the salient content of the shot [49]. The dynamic nature of video skims, which preserve sounds and motion, makes the summary entertaining and expressive [50]. On the other hand, the second summarisation technique, which consists in displaying a selection of representative frames, allows to create summaries in a more rapid and compact way [47]. The work of Smith and Kanade [49] describes a way to produce video skims performing four main actions: TF-IDF (Term Frequency Inverse Document Frequency) on the audio transcript to identify the important words, scene segmentation using histogram difference analysis, camera motion analysis to exclude scenes with excessive camera motion, object detection to identify human faces and text in the frames. The result is a combination of scenes with little or no motion, whose transcript contains a relevant word, or where faces or text are detected. Already in 1995, this rudimentary method proved the power of the integration of speech, language and image information.

An example of key frame-based video representation is given by the research of Ciocca and Schettini [47]. In 2006, the authors experimented a frame selection technique based on maximizing the difference measure between frames within each shot, described by color histogram, wavelet statistics and edge direction histogram. The advantage of this method over others, such as clustering strategies, is its limited computational complexity, that allows to extract key frames on the fly, even without processing the whole shot. In addition,

it can be used on compressed videos, too, avoiding the need to decode the frames.

### 2.2.3. OVERVIEW OF RECENT SUPERVISED AND UNSUPERVISED METHODS

A more recent work on static video summaries was conducted by De Avila, Sandra Eliza Fontes, et al., who developed the Video SUMMARization (VSUMM) approach [50]. In this case, the key frames correspond to the centroids of frames clusters computed using the k-means algorithm. Having to face the problem of lack of a consistent evaluation framework, the authors realised a new evaluation methodology, called Comparison of User Summaries (CUS), which consists of comparing automatic generated summaries to the ones manually created by users, considered as ground-truth. The performance was assessed by introducing two metrics: the accuracy rate and the error rate, respectively the percentage of matching and non-matching key frames. The method was tested on videos from the Open Video Project (OVP)<sup>4</sup> and the results outperformed previous techniques, such as DT (Mundur et al., 2006) [29] and STIMO (Furini et al., 2010) [51] and OV (DeMenthon et al., 1998) [51].

Among the methods that have been mentioned, VSUMM is one of those that rely on fully unsupervised settings. However, Potapov, Danila, et al. claim that in large video collections it is convenient to identify clusters of typical categories, for example birthday parties and flash mobs, that are useful to exploit recurrent visual content and repeating patterns [52]. They follow a category-specific summarisation approach, the Kernel Video Summarisation (KVS), characterized by two important components: a kernel temporal segmentation (KTS) algorithm, capable to identify general change points, and an importance scoring algorithm, to highlight the segments that are to be included in the summary. The first component is able to statistically discriminate change points in a few operations: the Gram matrix, which contains the frame-to-frame similarities, is computed from the frames visual descriptors, then the matrix is used in a linear programming problem to agglomerate similar frames, eventually leading to find the segments end points. The second algorithm is based on binary SVM classifiers, each trained for one specific category exploiting human annotations on the whole videos, that take as input a video segment and predicts if it is relevant. Finally, the most important segments are selected until a preset summary duration is reached, and then concatenated.

In parallel with KVS, in 2014 Grauman and Sha [53] experimented a different supervised approach, training a subset selection system using human-created summaries and the respective original source videos, from YouTube, the Open Video Project<sup>5</sup>, and the Kodak consumer video dataset [54]. Their technique is based on a probabilistic model for subset selection, called sequential determinantal point process (seqDPP), used for the extraction of relevant frames or subshots. The problem of dealing with a human-created training set is that different individuals might compose different summaries. To overcome it, the authors synthesize an "oracle" summary per video, which is the most similar to all user-generated ones in terms of Precision, Recall and F-score. The same metrics are used to assess the results of the seqDPP model, which prove the system superior than former supervised methods like VSUMM [50], due to its ability of acknowledging distant frames diversity in spite of the visual content.

The challenge of supervised learning video summarisation is undertaken, two years later, by Zhang, Ke, et al. [45], who relied on the efficacy of modelling variable-range temporal dependency of Long Short-Term Memory (LSTM) networks. This architecture makes sure that frames that are distant in time are distinguished even though the image descriptors are similar. In fact, the recurrence of the same scene in more parts of a video might be crucial to understand the content. The deployed neural network is a variant of LSTMs, called vsLSTM. The model is fed with a sequence of visual features in the input layer, which is followed by two LSTM layers, one to model the sequence in the forward direction and one in the backward direction. Finally, a multi-layer perceptron (MLP) outputs the probability of whether each element of the sequence should be included in the summary.

Following after Grauman and Sha [53], the model is further improved adding the determinantal point process (DPP) on top of the output layer. This combination allows to overcome the flaws of LSTMs and DPP that generally perform, respectively, high in recall and lower in precision and high in precision and lower in recall. The results confirmed the ability of dppLSTM to capture sequential semantic, even though temporal dependencies are captured better in videos where the content changes smoothly. In spite of the encouraging performance, the drawback of LSTMs of needing a large amount of annotated data for training still affects their versatility. Figures 2.1a and 2.1b show a graphical representation of the two architectures engineered by Zhang, Ke, et al. [45]

<sup>4</sup>Available at <https://open-video.org>

<sup>5</sup>Available at <https://open-video.org>

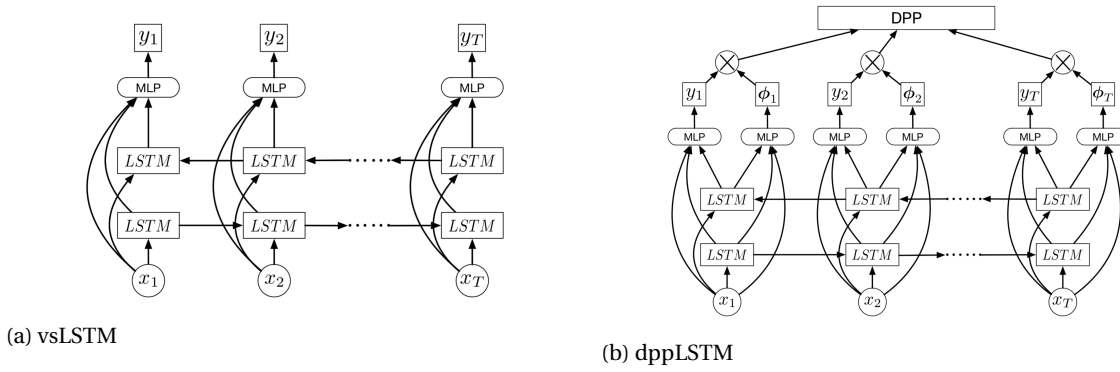


Figure 2.1: LSTM-based models used by Zhang, Ke, et al. [45] to predict the probability a frame or subshot should be included in the video summary.

In 2016, a group of researchers from Microsoft develop a technique for the summarisation of first-person videos [27], namely movies depicting a single individual, recorded via wearable devices, such as GoPro cameras and Google Glass. Such videos are usually long, redundant and unstructured. The deployed system is based on two-stream deep convolutional neural networks (DCNN), in order to model both spatial and temporal information. More specifically, the architecture consists of the combination of AlexNet [55] and C3D [56]. The model takes as input the original video, previously segmented, and produces a highlight score that reflects the importance of each subshot. Subsequently, the most relevant clips are selected to be part of the summary. In order to train the system of DCNNs to produce correct scores, a pairwise ranking layer is added on top. The network is trained using sets of labelled pairs of shots, one to highlight and one to discard: if it fails to rank the two segments in the right order, the gradient is backpropagated to the lower layers, whose parameters are adjusted accordingly.

The final summary can be a video skimming (HD-VS) or a video timelapse (HD-VT), in which all the shots are shown at lower or higher speed rate, depending on their highlight score. The authors used human evaluation to compare their results to previous techniques. The summaries were to be assessed in terms of coverage (which summary reports more accurately the content of the original video?) and presentation (which summary presents better the content of the original video?). The users showed a strong preference for highlight-driven summaries, both HD-VS and HD-VT, over uniformly sampled subshots, KVS importance-driven summary [52] and interestingness-driven summary [57].

#### 2.2.4. STATE-OF-THE-ART

The articles describing the following techniques were published during 2018 and represent some of the latest advances in video summarisation.

The deep summarisation network (DSN) proposed by Zhou, Kaiyang, Yu Qiao, and Tao Xiang [58] paves the way for state-of-the-art unsupervised methods for video summarisation. The DSN is composed of an encoder and a decoder (Figure 2.2) The former is a convolutional neural network (CNN) that extracts the visual features from the frames and passes it to the second component, a bidirectional recurrent neural network (BiRNN) based on LSTM, which decodes the visual features and at the same time encapsulates information concerning the previous and next frames. The DSN learns to produce a probability score for each video frame. In this work, the authors introduced an innovative framework based on reinforcement learning, for training the decoder in an end-to-end fashion. The reward is given by the sum of two components, the first representing the result of a function measuring the dissimilarity (D) between frames close in time, and the second a degree of representativeness (R), formulated considering the video summary as an instance of the k-medoids problem [59].

Using the datasets SumMe [57] and TVSum [60], the summaries produced by the DR-DSN were tested against the results of previous techniques, including [45, 50, 57], in terms of diversity and representativeness. The outcome confirmed the method is superior to previous unsupervised techniques and comparable to contemporary supervised strategies. Interestingly, the reinforcement learning approach seems to imitate the human-learning process, in fact it reproduces the same importance scores as the ones learnt from human examples in a supervised fashion.

In spite of the great potential of this RL-based method, summarisation with a RL framework was fur-

ther improved by Zhou, Kaiyang, Tao Xiang, and Andrea Cavallaro [32]. The authors introduced a weakly-supervised RL framework, that is able to build customized summaries for specific video categories, only using video-level category labels, which are easy to obtain. This framework contains a summarisation network with deep Q-learning (DQSN), which sequentially removes frames from a given video, depending on the prediction of the future rewards. Then, a classification network produces a global recognisability reward, which assesses whether the summary generated by the DQSN contains sufficient information to be classified to the original video category. Therefore, the summary is constructed in a way it is guaranteed to capture category-specific information. In addition, the classification network calculates the so called local relative importance reward, which, at any stage of the summarisation process, gives feedback to the partially generated summary. This dense reward is used to mitigate the effects of the credit assignment problem, which is a well-known issue of RL and expresses the difficulty in associating a certain action with its reward. The advantages of this architecture are the computationally efficiency, the need of no specific annotations, the state-of-the-art summarisation results on TvSum [60] and CoSum [61].

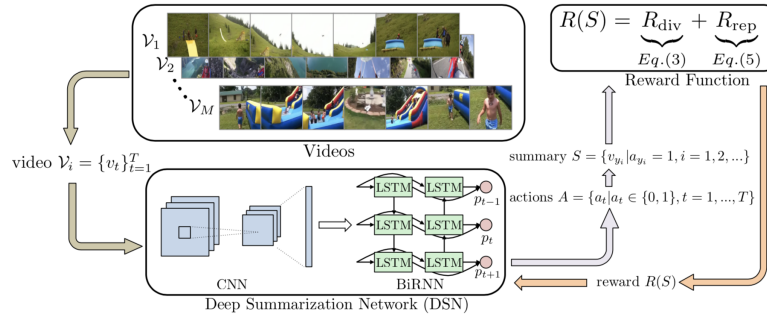


Figure 2.2: Deep summarisation network (DSN) by Zhou, Kaiyang, Yu Qiao, and Tao Xiang [58]

Subsequently to [58], Zhang, Yujia, et al. exploit generative adversarial networks to build a state-of-the-art architecture [62], that outperforms the DSN, obtaining 2.5% and 1.0% higher F-measure, tested on SumMe [57] and TVSum [60]. The contribution made by the authors is a new GAN model, built upon Dilated Temporal Relational (DTR) units, which overcome the shortcoming of the commonly used LSTMs of overlooking long-range temporal dependencies. More specifically, a DTR unit performs dilated convolutions across the temporal dimension, to capture temporal relational dependencies among video frames at multiple time scales. The generator, used in the system to predict the key frames, is composed of three modules. First, a temporal encoding module made of a bidirectional LSTM (Bi-LSTM) network and a network containing three DTR layers, receives as input the full length frames sequence and produces two sets of enhanced features for all frames. Secondly, these two types of features are assembled together by a compact video representation module. Finally, a summary predictor, that uses a fully-connected layer, a dropout layer and a batch normalization layer outputs one score for each input frame. On the other hand, the discriminator is used in the adversarial learning phase to distinguish among ground truth summaries, generated summaries, and summary consisting of random frames. It relies on a Bi-LSTM network for each input (full video or summaries), followed by three dense layers and a sigmoid non-linearity, to produce the discriminator score. The two models are jointly trained until the generator is able to fool the optimized discriminator. Then the generator is ready to create summaries in the inference process. The results proved that the frames representations learnt deploying the multi-scale temporal convolutions of DTR units help create complete but compact summaries, better than the previous state-of-the-art.

Most of the automatic summarisers found in the literature, including [58, 62], are designed following the same pattern: the video is represented in form of sequential data (frames or subshots) that are processed and labelled by a certain variant of recurrent neural networks, like LSTMs. Rochan, Mrigank, Linwei Ye, and Yang Wang [46] explore a different approach, completely based on CNNs, which allow to parallelise the elaboration of the input data, thus being able to exploit the power of GPUs and reduce the computational time. It consists in rerouting the problem of summarisation to semantic segmentation, using different input dimensions and number of channels. To this extent, instead of assigning semantic labels to image pixels, the classifier takes as input visual descriptors for the entire video and outputs one binary importance score for each frame.

A popular segmentation model such as FCN [63] is the base of the technique introduced by the authors, called SUM-FCN, consisting of temporal convolutions, temporal pooling, and temporal deconvolutions, in-



cluding skip connections between layers. The model is trained on the combination of the datasets SumMe [57], TVSum [60], YouTube and OVP [50]. In addition, an unsupervised variant of the model is proposed, where the network is optimized to reduce the pairwise similarity between the selected key frames. After testing against other current techniques, the competitive results of SUM-FCN proved CNNs can be used as a worthy replacement for LSTMs.

### 2.2.5. DISCUSSION

In the previous sections, an overview of the existing methods for automatic video summarisation was presented. Primitive techniques mainly focus on clustering visually similar frames under the same group and select one representative frame from each group [47, 49, 50], without any domain knowledge about the video category or its content. More recent works are mainly based on neural network, which are deployed to classify relevant frames or shots, using, for example, importance scores [52] or highlight scores [27], or are used to calculate the rewards in RL frameworks [32, 58]. As it was described in the introduction, the subject of the thesis is an automatic system capable for summarizing videos corresponding to type I (videos where one person speaks facing the camera, see Figure 1.1). Having to deal with such a static setting, it is convenient to a supervised learning approach, taking advantage of the recurrent common features of the video. This kind of approach was inspired by [45, 52, 53].

As can be observed in many works in the literature, several summarisation systems share a common structure: first the video is split into phrases or subshots, that are fed to a classification model trained to produce a score that expresses the probability of the frame or subshot in question to be included in the video summary. This inspired the idea of starting the highlights extraction pipeline (see Figure 1.2) with video segmentation [52], followed by the "saliency" score prediction. Different examples for the classification can be found in the literature, among which KVS importance-scoring algorithm [52], interestingness-driven summary [57] or highlight detection [27].

The evaluation of the results is fundamental to assess the final performance of the system. Many supervised methods in the literature have been assessed by comparing to ground examples, for example SumMe [57] and TVSum [60], in terms of precision/recall. Unfortunately, no datasets have ever been created for the particular case of video type I. Because of this, automatic evaluation is not possible. However, similarly to the case of the Comparison of User Summaries (CUS) introduced in [50], humans are going to be involved for the evaluation of the results of this thesis.

## 2.3. VIDEO SEGMENTATION

The preliminary stage in the process of highlights extraction from videos where one person speaks facing the camera is the temporal segmentation of the input video. In fact, the classification models proposed in this thesis operate at a segment-wise level, predicting a "saliency score" for each analysed video fragment, which expresses its relevance within the full-length video. As discussed in §2.2, a simple method to achieve temporal segmentation is by including similar sequential frames in the same video fragment, in such a way that the difference between frames of the same clusters is minimized, while the difference between frames of different clusters is maximized. Such a technique is the base of Kernel Temporal Segmentation [52]. However, this is not the only technique that was considered. In fact, as from videos three channels of information can be extracted — text, audio and video — other options for segmentation might be based on audio or textual content. The following sections will give an overview of the most common approaches for video segmentation.

### 2.3.1. METHODOLOGY TO SEARCH FOR PAPERS

The papers that involve the subject of video segmentation were found through a search on Google Scholar [43], using the keywords: "video segmentation", "shot boundary detection problem", "kernel temporal segmentation".

### 2.3.2. KERNEL TEMPORAL SEGMENTATION

The Kernel Temporal Segmentation (KTS) [52] is a kernel-based change point detection algorithm, more resistant to noise than traditional scene change detection methods such as the one introduced in [64]. In fact, it performs segmentation employing a more general statistical framework that understands if abrupt changes in the visual content are actually due to changes in the scenes. The algorithm seeks for change points from a positive-definite kernel matrix, which describes the frame-to-frame similarities, in terms of image descriptors. It can be implemented adopting dynamical programming. Adding a penalty term in the objective func-

tion prevents the algorithm to increase the intra-segments similarities by detecting too many change points. This allows to find the optimal number of change points, thus avoiding over or under-segmentation.

### 2.3.3. SHOT BOUNDARY DETECTION PROBLEM

Some works address the problem of video segmentation from the perspective of detecting when a shot starts and ends. This can be summarised as the "shot boundary detection problem", which consists in recognising cuts, dissolves and fades, which are techniques used to transit from one shot to another. L. Baraldi, C. Grana and R. Cucchiara [65] point out the importance of developing a technique that balances the trade off between accuracy and speed. The authors introduce a method based on an extended distance measure that is capable of detecting effectively both abrupt and gradual. One other important contribution is the collection and publication of the RAI dataset, which contains broadcasting videos with manually annotated shots and scenes<sup>6</sup>.

Subsequently, M. Gygli proposes a "ridiculously fast" algorithm for [66] for the identification of shot boundaries based on Convolutional Neural Networks. The utilisation of layers that are convolutional in time allows to parallelise the prediction of change points from the same sequence of frames fed to the network. This results in an increase of speed never achieved before.

### 2.3.4. AUDIO SEGMENTATION

Although color histogram differences and motions between video frames are the most common techniques for the detection of change points, solely visual information does not guarantee that video segmentation results in the identification of semantic scenes. In [67], the authors prove that both audio and visual boundaries can be combined to obtain a final segmentation. The audio is first analysed by discriminating speech from non-speech class using a K-Nearest Neighbor classifier. The latter is further classified into music, environment sound and silence segments. The speech is then split employing a speaker change detection technique based on line spectrum pair (LSP) divergence shape analysis [68]. The boundaries predicted using the audio analysis are selected as the final change points of the segmentation if they are also chosen by the color correlation analysis. The experiments showed that audio break information can improve significantly the performance of video segmentation, especially within the case of TV news broadcasts, on which the method was tested.

The efficacy of audio segmentation applied to data from the broadcast news domains was further demonstrated by the techniques presented in [69]. These were experimented on a collection of news datasets consisting of the Albayzín-2010 Audio Segmentation Evaluation [70], the Aragón Radio database from the Corporación Aragonesa de Radio y Televisión (CARTV) [69] and environmental sounds from Freesound.org [71] and HuCorpus [72]. This collection was created in spite of the fact that including different data sources introduce more variability, which leads to an increased difficulty of the problem.

### 2.3.5. ONLINE CHANGE POINT DETECTION

A different type of segmentation than the ones previously mentioned, is the one designed and employed for audio sequences. The authors of [73] develop a system based on a Hidden Markov Model framework with finite state space. By substituting the forward smoothing recursion, previously employed in [74], with incremental optimization inspired by the incremental EM algorithm of [75], they managed to develop an online algorithm for audio segmentation. This technique works in a way to detect regions in audio streams where the same statistical properties are maintained, without exploiting any specific content-based prior knowledge. Such algorithm can be used both for musical notes and on acoustic scenes: it is interesting to point out that the method was successfully tested on scenes from the Office Live Dataset [76], proving that it is applicable also to film productions, therefore possibly relevant for automatic video analysis tasks.

In 2018, another research on online change point detection proved that is possible to apply a parametric model, such as the Exponential Weighted Moving Average (EWMA) algorithm, without any prior knowledge [77]. The EWMA has the advantage of being very fast, but it is usually non applicable in model-free situations. This can be avoided with the introduction of the "No-prior-knowledge" EWMA (NEWMA), that is based on the comparison of two EWMA statistics, corresponding to pools of recent and old samples. This way, the algorithm does not need prior knowledge nor to store any information during the computation, thus making it very suitable for online applications or on portable devices.

---

<sup>6</sup><http://imagelab.ing.unimore.it>

### 2.3.6. DISCUSSION

Even though several techniques for video segmentation exist, the most common rely on finding visual differences in the frames that compose the videos, such as in scene change detection or the shot boundary detection problem. However, these methods are not really applicable to the videos analysed in this thesis. In fact, videos in which one person speaks facing the camera are expected to be characterised by a rather static visual content. For this reason, segmentation methods based on sequencing the speech of speech transcripts might be a more suitable option. In fact, the videos contained in the Political Speeches Dataset introduced in this thesis were divided into segments according to the pauses in the speech. This allows to obtain video fragments in which the speaker pronounces fairly complete sentences.

## 2.4. MULTIMODAL LEARNING AND ITS APPLICATIONS

The McGurk effect [78] demonstrates how human verbal communication is enhanced by the presence of visual information. For example, lip reading can help understanding a speech even in noisy settings. Just like multimodal information supports quicker human comprehension of the surrounding environment and conversations with others, multimodal sources of information can be used in Machine Learning for a richer feature representation, that, in many extent, resulted to improve the accuracy in classification problems. This section presents some of the recent publication on multimodal Machine Learning, that inspired the methodology of this thesis.

### 2.4.1. METHODOLOGY TO SEARCH FOR PAPERS

The papers discussed in this sections were retrieved from Google Scholar [43] using the keywords "multimodal learning", "multimodal Machine Learning", "multimodal Deep Learning", as well as "multimodal sentiment analysis", "emotion recognition", "social signal processing". More papers were found searching among the references and the referrers of the papers previously encountered.

### 2.4.2. EXAMPLES OF MULTIMODAL MACHINE LEARNING

An example of multimodal learning is given by [79], in which the authors address the problem of automatic speech recognition from a combination of audio and visual information. The visual features employed in the research are extracted from a mouth carving model using modules available on OpenCV, based on the ENCARA2 model [80]. As Deep Learning was proved high performing on both audio and visual classification tasks, multimodal learning is tackled with a bilinear bimodal Deep Neural Network (DNN). This technique consists in the fusion of two DNNs, one for audio input and one for visual input. The two DNNs are first trained separately, then the two category of processed data fused together by concatenating the outputs of final hidden layers. Surprisingly, the joint use of two information channels improve the classification performance both in presence of noise and in clean acoustic conditions.

MovieQA [81] provides another example where multimodal learning had a successful outcome. In this case, it is applied to the challenge of automatic story comprehension. The problem is addressed through the case study of developing a question-answering system about movie scenes with high semantic diversity. The authors extended the problem of scene understanding using a dataset of snapshots from movies, rather than just static pictures, as in former research on image caption generation and visual question answering [82–84]. To do so, they collected a new dataset containing video clips from 408 subtitled movies, enriched with the extended summaries from Wikipedia and the scenes transcripts. The question and answers annotations were collected from hourly-paid professional annotators. The question-answering system proposed in this research uses a multimodal feature embedding, which rely on mean pooled Word2Vec [85] for sentence representation, SkipThoughts [86] to model the sentence semantics, GoogLeNet [87] and hybrid-CNN [42] for frame-wise features, mean pooled over all frames in a shot. The latter is combined with word embeddings through a linear mapping. The results show that the fusion of these feature modalities outperforms the respective unimodal cases. The joint-use of textual and visual information can result in application also in retrieval systems, as in the case of image search tackled in [88], harnessing the common presence of captions and tags in addition to the image content. Other applications include image-to-image translation [5, 9], fusing multimodal domain-invariant and domain-specific properties.

Even in the topic of highlights extraction from videos it is possible to find examples of multimodal approaches. Recently, Merler, Michele, et al. [6] published a research on sport highlights detection. Their approach, called High-Five and based on a combination of information regarding the players' reactions and expressions, crowd cheering, statements from the commentator and game analytics, to determine the most

interesting moments of a game. Each feature modality is analysed independently in order to predict an "excitement score" for each video scene. The scores are rescaled in a range between 0 and 1 via sigmoid normalization and, finally, the total score is computed as a their weighted linear sum. The highlights are then composed from high scoring video segments, with excitement score above a predefined threshold. A similar research presented the newer SCSampler [89], which provides a system for "saliency detection" in videos, the same problem tackled in this thesis. Once more, this method is applied to sport videos, but at a clip level. Two distinct neural architecture assign audio and video saliency scores, which are finally combined to obtain an aggregated score, used to predict the most informative or exciting clips from the full videos.

As this multitude of research shows, the fusion of multimodal data can be beneficial for several applications, including multimodal highlights extraction. Since this thesis has the objective of extracting information from videos where one person speaks facing the camera, the literature focuses on the particular case of multimodal sentiment analysis and emotion recognition, which usually involves data of the same form as the one that is in our interest.

### 2.4.3. MULTIMODAL LEARNING IN SOCIAL SIGNAL PROCESSING

Social signal processing [90] is an emerging research field that studies how to classify and interpret social signals and social behaviours, like turn taking, politeness, and disagreement, in order to obtain a thorough understanding of the way human communicate with each other. The possibility to categorise human social signals allows to develop Artificial Intelligent technologies that can automatically recognize socially relevant information, e.g. the role of a person in a group, the detection of emotions or psychological states. Behavioural cues are fundamental in human communication and they are expressed by a combination of speech signals, body gestures, facial, and vocal expressions. The intrinsic multimodal nature of social signals gave rise to a growing interest by the social signal processing community on multimodal Machine Learning.

Emotion recognition is one of the topic involved in social signal processing. According to psychologists, humans transmit their emotions through the use of specific facial expressions. For this reason, automatic facial expression recognition (FER) is an important problem that can have applications in areas such as human-computer interaction [91]. Generally, the models developed for FER from visual data follow two general approaches: the first method utilises texture-related feature collected at a pixel level, while the second method is based on landmark, namely indicators for face key points whose movements can indicate the presence of a facial expression [92]. Whereas with the first approach it is possible to obtain very precise representations, but is very sensitive to image variations, like luminance and masking effect, the second cannot distinguish subtle changes in the visual content. For this reason, a multimodal approach based on the integration of the two can be beneficial to the classification performance. The adoption of autoencoders architectures with structured regularization can be used to learn automatically a joint representation of these two features modalities, by identifying the correlations and interactions between texture and landmarks [3].

To address the Emotion Recognition in the Wild Challenge (EmotiW)<sup>7</sup>, in 2013 Kahou, Samira Ebrahimi, et al. [4?] developed a multimodal classification model, based on the fusion of several Machine Learning architecture, each tailored for on one modality. These comprise a convolutional neural network for visual information from faces, a deep belief net for audio stream representation, a k-Means model to extract descriptors for the mouths expressions and a relational autoencoder, to model actions in the videos. These were all combined into a common classifier, by concatenating the probability vectors output by each model into fixed-length vectors, used to train a support vector machine (SVM) classifier, as well as a multilayer perceptron (MLP). The authors applied their method to the Acted Facial Expression in the Wild (AFEW) dataset<sup>8</sup>, containing video clips from feature-length film, of a duration of a couple of seconds, where the purpose was to recognise the seven emotions angry, disgust, fear, happy, neutral, sad, surprise. They won the EmotiW challenge in 2013, achieving a final accuracy of 35.58% only using ConvNets trained on one modality, and 41.03% with the proposed aggregated technique. The intrinsic complexity of the problem, together with variation among subjects, lighting and poses, makes it very hard to recognise emotions with high precision and recall.

In 2019, the EmotiW challenge focused on the task of predicting the level of engagement of a student, according to a four-level scale. The utilised dataset involves video records of students watching online courses, divided into segments. The winning team [93] based their approach on the joint usage of features describing the movement of the head, the gaze and the body posture, which were extracted with OpenFace and OpenPose [94, 95]. In their method, the different types of features are processed independently by a 1 or 2 layer

<sup>7</sup><https://icmi.acm.org/2019/index.php?id=challenges>

<sup>8</sup>Available at <https://computervisiononline.com/dataset/1105138659>

LSTM network and three fully connected layers. The outputs are combined together through an ensemble technique, giving to all the feature modalities the same weights. This approach resulted into the lowest MSE achieved in the competition, namely 0.0626.

In 2018, Liu, Zhun, et al. [96] proposed a unified method, namely the Low-rank Multimodal Fusion, to integrate multimodal features for different application, including sentiment analysis, speaker trait analysis, and emotion recognition. As one of the main difficulties in multimodal learning is the increase in computational complexity due to the great dimensionality of the input data, Low-rank Multimodal Fusion has the advantage of drastically reducing computational costs. In fact, multimodal low-rank weight tensors can be decomposed in parallel and combined together in order to achieve compact multimodal representation. For this reason, tensors are a good option to represent intra-model and inter-modal dynamics.

The integration of multimodal features can be performed through a parametric function, whose parameters are learnt during training. This is the option developed for the Local-Global Ranking Fusion (MLRF) method [8], where an LSTM network is trained to capture temporal dependencies across multimodal time series data and predict the multimodal feature vector. The same authors of the Local-Global Ranking Fusion (MLRF), experimented several techniques for multimodal fusion. One of these is the Memory Fusion Network [12], which is replicated in this thesis and is described into more details in the next section.

#### 2.4.4. MEMORY FUSION NETWORK

In the paper "Memory Fusion Network for Multi-view Sequential Learning" [12], the authors deal with the problem of multi-view sequential learning. This is defined as the natural extension of multi-view learning, which is the problem of dealing with information coming from different input modalities, to the case of sequential data. In this setting, they develop a model able to identify two forms of interactions in the data: "view-specific interactions", that are the temporal relations within a single feature modality, and "cross-view interactions", that describe relations across different views. This work outperforms the significant drawbacks of multimodal learning methods based on the naive concatenation of multimodal features: the higher risk of overfitting and the incapacity to model the unique statistical properties of a single modality. In addition, it is superior to approaches that consist in collapsing the time dimension to reduce the problem to a generic multi-view learning problem, as the former takes into account temporal relations.

The proposed architecture, the Memory Fusion Network (MFN), is composed of three main parts: a system of LSTMs that process the input modalities separately, the Delta-memory Attention Network (DMAN) that seeks for correlations across memories of different LSTMs, and the Multi-view Gated Memory that stores the cross-view interactions over time.

In the System of LSTMs, each neural network  $n$  follows the following update rules at every time step  $t$ :

$$\begin{aligned} i_n^t &= \sigma(W_n^i x_n^t = U_n^i h_n^{t-1} + b_n^i), \\ f_n^t &= \sigma(W_n^f x_n^t = U_n^f h_n^{t-1} + b_n^f), \\ o_n^t &= \sigma(W_n^o x_n^t = U_n^o h_n^{t-1} + b_n^o), \\ m_n^t &= W_n^m x_n^t = U_n^m h_n^{t-1} + b_n^m, \\ c_n^t &= f_n^t \odot c_n^{t-1} + i_n^t \odot m_n^t, \\ h_n^t &= o_n^t \odot \tanh(c_n^t). \end{aligned}$$

In the equations,  $i$ ,  $f$  and  $o$  refer to the input gate, forget gate and output gate of the LSTM cells, which, together with the proposed memory update  $m$ , contribute to modify the inner memory  $c$  and define the network output  $h$ .

For what concerns the DMAN, a neural networks receive as input the concatenation of the memories  $c_n$  at time  $t$  and  $t - 1$  and calculates softmax activated attention coefficients,

$$a^{[t-1,t]} = \mathcal{D}_a(c^{[t-1,t]}),$$

which are used to derive the attended memory of the LSTMS:

$$\hat{c}^{[t-1,t]} = c^{[t-1,t]} \odot a^{[t-1,t]}.$$

In  $\hat{c}^{[t-1,t]}$ , the elements of  $c^{[t-1,t]}$  that contribute in a cross-view interaction are amplified. Since the tensors  $c$  are memories, also cross-view interactions that happened before  $[t - 1, t]$  can be detected. The two gates  $\gamma_1$

and  $\gamma_2$  that control the Multi-view Gated Memory use the update proposal  $\hat{u}^t$ , defined as

$$\hat{u}^t = \mathcal{D}_u(\hat{c}^{[t-1,t]}),$$

to update the memory  $u$  that describes the history of cross-view interactions:

$$u^t = \gamma_1^t \odot u^{t-1} + \gamma_2^t \odot \tanh(\hat{u}^t).$$

Finally, the output of the system of LSTMs  $h_n$  and the final memory state  $u$  of the Multi-view Gated Memory are concatenated and used jointly to calculate the final one-dimensional prediction.

The MFN architecture is used by the authors for the task of multimodal sentiment analysis, which is based on the joint usage of audio, textual and visual descriptors to predict a positive or negative sentiment score, ranging from -3 to +3. The model was tested on different datasets, all containing mostly opinion videos or video reviews, among which the MOSI dataset [38], the MOUD dataset [2] and the Youtube dataset [1]. The model was benchmarked against previous view concatenation and multi-view sequential learning models, like [97] and [98], and was confirmed as state-of-the-art for multi-view sequential learning.

### 2.4.5. DISCUSSION

In this section, several research concerning the subject of multimodal Machine Learning and techniques for multimodal feature integration were presented. Among the multiple application of multimodal learning, which comprise speech recognition, multimedia content indexing and retrieval, and effective computing and media description [99], the literature review focuses on emotion recognition. This interest is motivated by the fact that experiments on this topic usually involve video data, where the subject is a person filmed by a camera. These videos are very similar to the ones analysed in this thesis for the purpose of the Avengers Project (see Figure 1.1). The work conducted by the authors of the MOSI and MOSEI dataset [7, 8, 12, 37, 96] caught the attention, as it provides inspiration about how to extract features from videos where one person speaks to the camera and interesting Deep Learning architectures, like the Memory Fusion Network. The idea of using MFN for the task of information extraction is carried out throughout this thesis project. The next section, presents some of the most famous multimodal datasets and the ones that inspired the creation of the "Political Speeches Dataset" first introduced in this thesis.

## 2.5. MULTIMODAL DATASETS

The main contribution of this thesis project is the introduction of the novel multimodal "Political Speeches Dataset". In order to collocate the dataset in the context of multimodal learning research and understand its relevance, the similarities and differences from the already existing multimodal datasets, it is worth to include this aspect of multimodal learning in the literature review. In this section, some of the most relevant multimodal datasets are introduced. Particular attention is drawn on the MOSI and MOSEI datasets, which directly inspired the creation of the "Political Speeches Dataset" and the feature extraction process.

### 2.5.1. METHODOLOGY TO SEARCH FOR PAPERS

Literature regarding multimodal datasets was encountered during the investigation of research on multimodal learning. Therefore, the search methodology corresponded to the one from §2.4.

### 2.5.2. MULTIMODAL DATASETS

The literature review came across several researches on multimodal Machine Learning, each addressing different objectives and application. For this reason, many multimodal datasets were encountered. Among the most interesting, it is worth to mention Microsoft COCO (MS COCO)<sup>9</sup> [100] and MIRFLICKR-1M<sup>10</sup> [101], very important for the task of image captioning. MS COCO contains over 330,000 images equipped with human generated captions, collected through crowdsourcing [102], whereas MIRFLICKR-1M includes up to 1M images from Flickr with user tags and EXIF metadata. Other datasets comprising a collection of textual and visual information are The Wikipedia and British Library<sup>11</sup> datasets [103, 104], which contain text chunks coupled with images, scraped from Wikipedia articles and books from the British Library in their digitalised version.

<sup>9</sup> Available at <http://cocodataset.org/#home>

<sup>10</sup> Available at <https://press.liacs.nl/mirflickr/>

<sup>11</sup> Available at <http://www.cs.cornell.edu/~jhessel/concreteness/concreteness.html>

For what concerns the social signal processing domain, the AMI Meeting Corpus<sup>12</sup> [105] is utilisable to study human conversational behaviours, as it is composed of 100 hours of meeting recordings with manual annotations, including word-timed speech transcriptions, dialogue acts, types of head gesture, hand gesture, and gaze direction, hands position, movements around the room and emotional states. On the other hand, the SEMAINE corpus<sup>13</sup> [34, 35] and the RECOLA multimodal database<sup>14</sup> [36] offer videos of people frontally recorded, while they are holding remote conversations through video call transcribed and annotated with regard to basic emotions, epistemic states and the interaction process.

### 2.5.3. THE MOSI AND MOSEI DATASETS

The datasets that are most relevant to this thesis project are the CMU Multimodal Opinion Sentiment Intensity (MOSI) dataset [12] and the CMU Multimodal Opinion Sentiment and Emotion Intensity (MOSEI) dataset [37]<sup>15</sup>, because of their similarity to the type of videos that are the target of this thesis. The MFM proposed by [12] is implemented to be used on the MOSI dataset [38], created by the same authors. The MOSI dataset is a collection of 89 opinion videos from YouTube, containing one person speaking in front of the camera. The subjects are 41 women and 48 men, approximately between twenty and thirty years old, all English speakers but from different ethnic backgrounds. From the videos, 3702 segments were extracted, among which 2199 opinion segments and 1503 objective segments. The dimension of the MOSI raw dataset is 3.6 GB, including full and segmented videos. Three modalities of features have been extracted from the segments.

**Language view** Manual speech transcription, aligned with the audio using P2FA<sup>16</sup> to obtain the timestamps. The words are then represented with GloVe [106], which result in 300-dimensional features.

**Visual view** Automatically extracted visual features, including 47-dimensional features from Emotient FACET 4.1 [107].

**Acoustic view** Audio 74-dimensional features including pitch, energy, NAQ (Normalized Amplitude Quotient), MFCCs (Mel-frequency Cepstral Coefficients), Peak Slope, Energy Slope extracted automatically with COVAREP [108].

The video segments are labelled based on the sentiment intensity of the subject speaking. The annotations were manually collected by workers from Amazon Mechanical Turk<sup>17</sup>. Eight possible values were allowed: strongly positive (+3), positive (+2), weakly positive (+1), neutral (0), weakly negative (-1), negative (-2), strongly negative (-3). Each video was annotated by five people and the final label is given by the average of the five scores.

The MOSEI dataset [37] is a larger version of MOSI [38], put together by the same authors. It contains a selection of 3228 videos from YouTube where one person is speaking in front of the camera. The features that describes the videos are collected in a similar way as MOSI.

**Language view** From the manual transcription of the speech, 300-dimensional Glove word embeddings [106].

**Visual view** 35-dimensional highlevel visual features describing the facial expressions, extracted with Emotient FACET 4.2 [107].

**Acoustic view** 74-dimensional acoustic features extracted with COVAREP [108].

All the videos are annotated with two type of labels.

- **Sentiment Labels:** similarly as MOSI, one label per video with value ranging between -3 and +3, indicating whether the sentiment is negative or positive.
- **Emotion Labels:** six values indicating the presence of the emotions happiness, sadness, anger, fear, disgust and surprise. The labels can have value from 0 (no evidence of the emotion) to 3 (highly present emotion).

<sup>12</sup>Available at <http://groups.inf.ed.ac.uk/ami/corpus/>

<sup>13</sup>Available at <http://www.semaine-db.eu>

<sup>14</sup>Available at <https://diuf.unifr.ch/main/diva/recola/index.html>

<sup>15</sup>Available at <https://www.amir-zadeh.com/datasets>

<sup>16</sup><https://github.com/ucbvlab/p2fa-vislab>

<sup>17</sup><https://www.mturk.com>

Dataset	Number of videos	Text features	Acoustic features	Visual features	Labels
<i>MOSI</i>	89	GloVe word vectors 300-dimensional	COVAREP 74-dimensional	Emotient FACET 4.1 47-dimensional	Opinion Labels [-3,+3]
<i>MOSEI</i>	3228	GloVe word vectors 300-dimensional	COVAREP 74-dimensional	Emotient FACET 4.2 35-dimensional	Opinion Labels [-3,+3] + Emotion labels [0,3] (happiness, sadness, anger, fear, disgust, surprise)

Table 2.1: Summary of MOSI and MOSEI characteristics.

#### 2.5.4. DISCUSSION

One of the main obstacles of multimodal learning is the lack of data, that spurs to the collection of application specific datasets with apposite annotations. The literature review and the discovery of the already existing multimodal datasets influenced the decision to collect the novel "Political Speeches Dataset". As the objective of this thesis is to extract information from videos where one person speaks to the camera, the MOSI and MOSEI datasets caught particular attention. Although the nature of the videos involved differ, the "Political Speeches Dataset" and the MOSI/MOSEI datasets share a similar structure. In fact, they both comprise collection of segmented videos, where one person speaks in front of a camera. The video fragments have similar duration and they contain the same types of features: Glove word embeddings [106], audio features from COVAREP [108] and facial descriptors, similar to the ones from Emotient FACET 4.2 [107] but extracted with OpenFace[94].

## 2.6. CROWDSOURCING

Wherever a high level understanding of data is needed, human knowledge can be exploited to provide support for Artificial Intelligence. For instance, humans can be involved for the purpose of collecting data and manually annotating them, which is fundamental for the creation of many datasets. Another possibility, is to ask humans to evaluate the results produced by Machine Learning algorithms, in order to assess their performance and understand what need to be improved. The recruitment of contributors and the task completion process is part of so-called "crowdsourcing". This section presents the subject of crowdsourcing and some of its applications, that inspired the evaluation method of this thesis.

### 2.6.1. METHODOLOGY TO SEARCH FOR PAPERS

The articles that are mentioned in this section were found on Google Scholar [43], using the keyword "crowdsourcing". Among the results, papers with the highest number of citations were considered. Additional papers are found through the research on automatic summarisation, as in many cases crowdsourcing is employed for data annotating or for the evaluation.

### 2.6.2. THE USE OF CROWDSOURCING IN SUPPORT OF MACHINE LEARNING

The term "crowdsourcing" was first coined in 2006 by Jeff Howe [109], who described it in relation to outsourcing, as an advancement that allows for cheap labour from everyday, non-professional people. In 2011, Enrique Estellés-Arolas and Fernando González-Ladrón-De-Guevara gathered forty different definitions of "crowdsourcing" from the existing literature [110]. To cite a few, "the mechanism by which talent and knowledge is matched to those who need it" [111], "a way of outsourcing to the crowd tasks of intellectual assets creation, often collaboratively, with the aim of having easier access to a wide variety of skills and experience." [112], "a new online distributed production model in which people collaborate and may be awarded to complete task." [113] and "focal entity's use of an enthusiastic crowd or loosely bound public to provide solutions to problems." [114].

In the context of Artificial Intelligence, crowdsourcing is often used to provide structured data for Machine Learning tasks, reliable annotations of to evaluate the results of research. In most cases, crowdsourcing tasks are implemented using online platforms, such as the popular Amazon's Mechanical Turk<sup>18</sup> [115, 116] and Figure Eight<sup>19</sup> (former CrowdFlower).

One of the largest database of images is ImageNet<sup>20</sup> [117]. The annotation process was performed via crowdsourcing using Amazon's Mechanical Turk, that allowed to obtain over 14 million labelled images in a

<sup>18</sup><https://www.mturk.com>

<sup>19</sup><https://www.figure-eight.com>

<sup>20</sup>Available at <http://www.image-net.org/index>



time span of approximately two years. In 2016, Sigurdsson, Gunnar A., et al. [118] used crowdsourcing to collect the new Charades dataset<sup>21</sup>, containing hundreds of video records of people conducting every day activities in their home. The video recording process, as well as the video annotation concerning action classification, localization, and video description was fully crowdsourced: the workers were asked to record themselves while executing the actions described in predetermined scripts, also collected through crowdsourcing, and to provide manual labels. The Charades dataset can be used to train systems for human activities recognizing in realistic home environments. Another example of crowdsourced dataset is given the corpus of facial expression presented in [119], consisting of 3,268 videos of people reacting to visual stimuli. Similarly, Mohammad Soleymani and Martha Larson [120] designed a crowdsourcing job to collect the 2010 Affect Task corpus for the prediction of viewer-reported boredom. Regarding information retrieval, [121] presents a news query classification dataset, composed of pairs of queries and news articles, gathered by asking crowdworkers to label queries as news-related or not.

One Machine Learning field in which crowdsourcing has proved to be an optimal solution, is in human-machine translation. In 2012, crowdsourcing was used to collect a parallel corpora between English and six languages from the Indian subcontinent [122]. This process worked was successful and the cost resulted much lower than if involving an expert, in spite of the high degree of morphological complexity of the Indian languages, their diversity from English and the low-resource of studies of these languages. The crowdsourcing tasks were designed in a way that each worker was asked to translate one word or one sentence extracted from Wikipedia articles, and, in order to prevent low-quality translations, the authors designed an additional control task to spot eventual mistakes. Other example concerning human-machine translation is [123], where the authors developed an automatic way to discern good and bad quality translations annotated by non-professionals workers.

The Chimera solution [124] proved that crowdsourcing works also at large scales, with the necessary attention. In this case, the authors tackled the problem of product classification at WalmartLabs, with objective of classifying tens of millions of product descriptions into more than 5000 product types. Due to the great variety of items that can follow under the same category, designing an effective crowdsourcing annotation task can be critical. In fact, it would be too difficult for a crowdworker to examine more than 5000 labels in order to find the right one. However, the authors managed to shift the use of crowdsourcing to the evaluation process, by first generating a set of hand-crafted expert rules, classifying the object, and asking the crowdworkers whether the classification was correct. This way, the correctly classified items can be used to learn the characteristics of those items. The present research provides an example in which crowdsourcing is used jointly with expert knowledge in a scalable way.

Crowdsourcing find application also in automatic video summarisation, which is tightly linked to highlights extraction. The SumMe benchmark [57], already mentioned in §2.2, comprises human created summaries from 25 videos covering holidays, events and sports, from 1 to 6 minutes long. The contributors were asked to manually compose a summary conveying the important information, by manually cutting and selecting the relevant scenes. SumMe was used by its authors for the evaluation of video summarisation methods, that is, automatic video summaries were compared to the ground truth collected summaries in terms of f-measures. Similarly, Khosla, Aditya, et al. [125] rely on crowdsourcing for the generation of ground truth summaries, used to evaluate the results of their unsupervised summarisation algorithm in terms of precision and recall.

Other application of crowdsourcing include relevance assessments [126], causality detection in narrative texts [127], medical images classification [128] and geospatial data collection [129].

### 2.6.3. DISCUSSION

In this section, some applications of crowdsourcing in Machine Learning were presented. Normally, crowdsourcing is used to extract organised manually annotated data from a mass of unstructured data, to generate the data themselves, or to evaluate the results of newly proposed techniques. In video summarisation, former research rely on crowdsourcing to generate ground truth summaries, used to assess the performance of supervised and unsupervised summarisation approaches [57, 125, 130].

In spite of the availability of the large MOSI and MOSEI datasets, which contain videos of people speaking in front of the camera (therefore optimal for the Avengers Project), the only annotations contained in these datasets regard sentiment intensity and emotions, which alone are not so much indicators of saliency and relevance. The lack of labelled data represent one of the main difficulties in the highlight extraction prob-

---

<sup>21</sup>Available at <https://allenai.org/plato/charades/>

lem tackled in this thesis. The option of using crowdsourcing to gather ground truth summaries, just like the cases previously mentioned, was taken into consideration. However, summary creation tasks are non trivial: in order to be able to select the relevant parts of a video, the worker has to watch it carefully and then manually select the important scenes. This is only possible with simple videos of limited duration, otherwise the identification of relevant information might be hard to execute by someone with no background knowledge about the content and no filmmaking skills. In addition, the scene cutting and selection process requires a lot of handcraft, which come at a cost. Similarly to [124], due to costs and time constraint, in addition to the intrinsic difficulty of manual summarisation/highlights detection from videos, this thesis is going to employ crowdsourcing to perform a human evaluation of the results. Further details about the use of crowdsourcing will be given in the next chapters, in particular Chapter 3, 6 and 7.

## 2.7. CONCLUSION

In this chapter, the most relevant research areas to this thesis project were introduced. Among these, particular attention was drawn on the topic of automatic video summarisation (see §2.2), which is analogous in many aspects to automatic highlights extraction. The review of the research on video summarisation influenced the design of the framework shown in §1.2, namely the decomposition of the problem into: video segmentation (discussed in §2.3), segment-wise video classification, segments selection and the composition of the final highlight clip. While [11] inspired the use of multimodal features for scene understanding and the methodology for scene clustering, the review of projects on multimodal learning (see §2.4) suggested the use of feature fusion techniques, such as the MFN architecture [12].

During the literature review, several multimodal datasets were discovered (§2.5); among these, the most relevant to this research are the MOSI [38] and MOSEI [37] datasets. Nevertheless, these datasets do not suffice to train Machine Learning models for highlights extraction from videos, as they do not provide annotations concerning saliency, necessary to detect the highlights in a supervised fashion. The lack of suitable "saliency" labels led to the investigation of crowdsourcing approaches for data annotating (see §2.6). However, since such an annotating task would be too complex and time consuming, crowdsourcing is going to be used for the evaluation of the results instead. The problem of saliency labels deficiency was avoided by developing a process for automatic data collection and data labelling, that focuses on videos of political speeches.

Further information about the methodology designed to answer the research questions and the novel dataset, collected to resolve the lack of labelled data, follows in chapters §3 and §4.

# 3

## METHODOLOGY AND MODELS FOR THE HIGHLIGHTS EXTRACTION

### 3.1. INTRODUCTION

The focus of this research project is to investigate how the task of highlights extraction from videos can be effectively automated through the utilisation of state-of-the-art techniques in Machine Learning.

The principal hypothesis upon which the thesis is based is that the power of the combination of multi-modal features can be used to effectively train Machine Learning or Deep Learning models that are capable to extract useful information from videos. In order to test the hypothesis, while following the purpose of the Avengers Project, the work focuses on the analysis of videos where one person speaks while being recorded by a front facing camera. The objective of highlights extraction of videos where one person speaks to the camera (type 1, see 1.1) provides the case study of the thesis.

The videos that are analysed in this thesis convey information through three distinct channels:

1. a textual channel that corresponds the content, namely what the speaker says in the video, which can be manually or automatically transcribed;
2. a visual channel, given by the the facial expressions, the gestures, the posture of the speakers, that can be used as descriptors for the state of mind and the emotions of the speaker;
3. an audio channel, which is represented by the the voice of the speaker, the tone of the speech and the vocal volume.

Given this setting, it is possible to extract three types of multimodal features, exploitable to describe content of the speech, the use of facial expressions or gestures and the variations in the tone of the voice.

Due to the impossibility to record a large collection of videos where one person speaks facing the camera, with a sufficient duration and an adequate number of speakers, in restricted time and with a limited amount of monetary resources, a dataset was gathered from already existing materials, available online. More precisely, the leading idea was to concentrate the work around the analysis of political speeches. The main benefits brought by this choice are the large availability of video contents shared on the internet, the good quality of the footages and the consistency of the speeches themselves, which make the task of large-scale automatic video analysis feasible. All the information concerning the collection of the novel "Political Speeches dataset" are explained in the next chapter, 4.

In order to test whether multimodal descriptors are likely to make strong predictions about the relevant segments in a video, multimodal features extracted from the "Political Speeches Dataset" are used to train several Machine Learning models, from clustering algorithms to a complex neural network architecture. To verify the validity of the multimodal approach, the results from these models are compared against the ones obtained with a baseline method based only on one feature modality, namely the speech content, extracted from the respective transcript. The methodology is inspired by the papers "Harnessing AI for Augmenting Creativity: Application to Movie Trailer Creation" [11] and "Memory fusion network for multi-view sequential learning" [12]. From the first paper, a former project from IBM Research<sup>1</sup>, the idea of tackling the challenge

<sup>1</sup>[https://researcher.watson.ibm.com/researcher/view\\_group\\_subpage.php?id=9482](https://researcher.watson.ibm.com/researcher/view_group_subpage.php?id=9482)

of extracting interesting content from videos was derived. In addition, the same scene clustering approach based on multimodal features is replicated in this thesis. The second paper, that deals with multimodal sentiment analysis, suggested to focus on people close ups and to include emotion recognition to the set of descriptors. Moreover, the Deep Learning model introduced by the authors, that is the Memory Fusion Network, is adapted and repurposed for the information extraction problem addressed in this thesis.

This chapter describes in detail all the steps that form the research methodology. The scheme in figure 3.1 gives an overview of the main components.

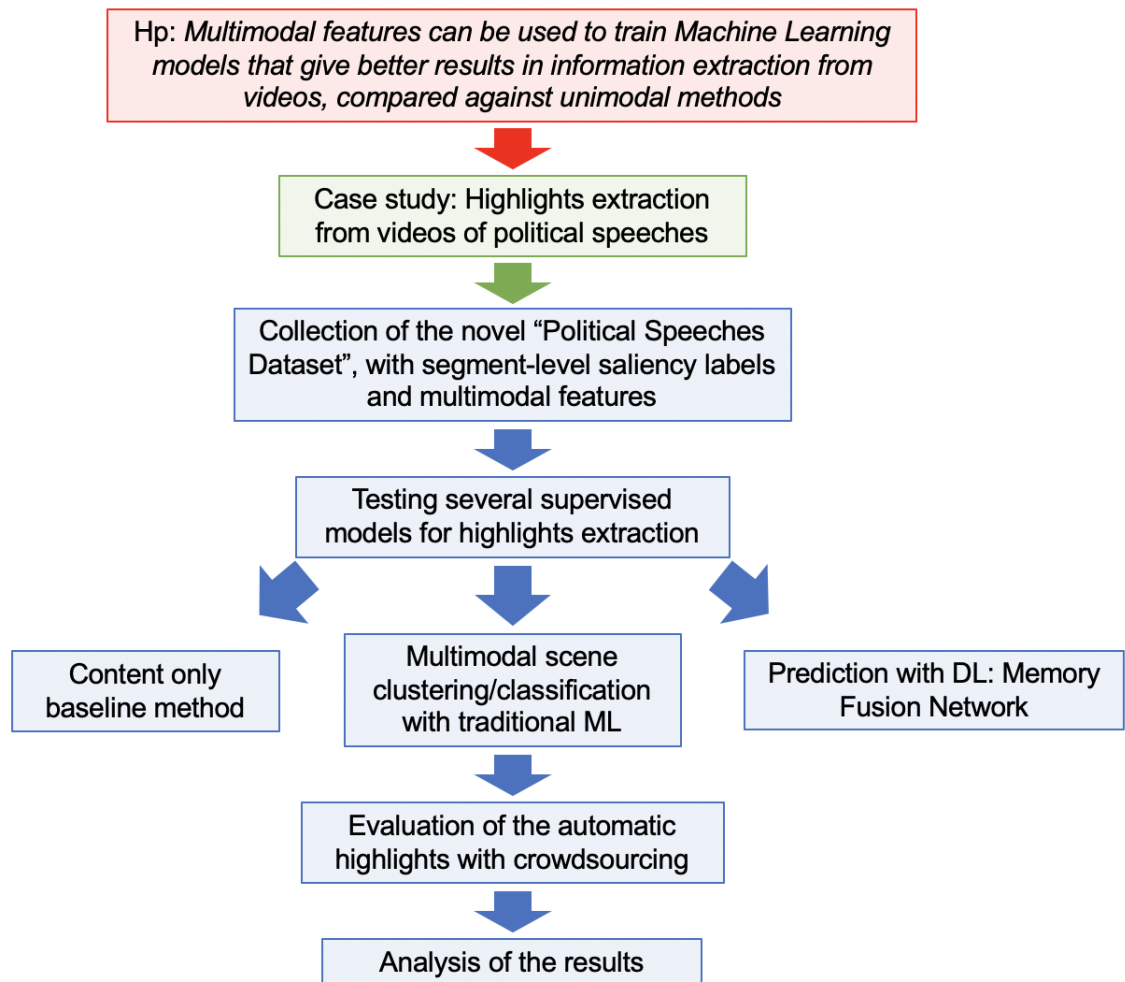


Figure 3.1: Methodology - Overview

### 3.2. DATASET

When dealing with automatic information extraction from videos, the first obstacle one has to face is the lack of datasets labelled appropriately. Despite the availability of several video datasets usable for different tasks<sup>2</sup>, a labelled dataset containing close ups videos of people does not exist. The closest examples are the MOSI and MOSEI datasets [37, 38], containing opinion videos from YouTube equipped with labels regarding the sentiment intensity and presence of emotions, but these are not directly usable for automatic highlights extraction. This aspect makes it very hard to approach the problem in a supervised fashion. On the other hand, unsupervised learning introduces additional difficulties for what concerns the evaluation of the results. In fact, without adequate labels it is not possible to estimate the accuracy of the results, or even train a model appositely for video analysis. Even though it is possible to perform human evaluation deploying

<sup>2</sup><https://www.di.ens.fr/~miech/datasetviz/>

crowdsourcing, the relative complexity and costs fall outside the resources available for this thesis projects. For the reasons mentioned, a novel dataset was collected appositely for the task of highlights extraction task from political speeches. The multimodal dataset collection constitutes the first step of the methodology and represents the biggest contribution of this thesis. The Political Speeches Dataset is used to train all the models that have been investigated and assess the results automatically through the dedicated "saliency labels". All the details about the motivation that led to the creation of the Political Speeches Dataset and the process to collect the speeches videos and multimodal features are reported in chapter 4.

### 3.3. MACHINE LEARNING METHODS FOR HIGHLIGHTS EXTRACTION

One common approach is proposed for the purpose of testing whether it is possible to use Machine Learning to perform the task of automatic highlights extraction from videos, objective of this thesis. Namely, a Machine Learning model is used to classify the videos in order to identify salient video segments. Then, the chosen video segments are concatenated, following their original temporal order, and the result is the highlight clip. If it can be demonstrated that the automatically generated highlight clips are of good quality, then it can be concluded that Machine Learning is able to analyse videos and extract highlights, that represent the salient content of the videos. This would answer **RQ2**.

To investigate the superiority of multimodal learning in the task of information extraction from videos, which is the object the subquestions of **RQ2**, several supervised Machine Learning models trained with multimodal features are tested against a purely content-based method, the baseline. The proposed Machine Learning approaches can be divided into three categories.

1. Unimodal classification: Baseline method
2. Multimodal features concatenation: Scene clustering/classification with traditional Machine Learning
3. Multimodal features integration: Saliency prediction with Deep Learning using MFNs [12]

While the first method is uniquely based on information extraction from the speeches transcripts, the other two classes of models are designed appositely to test the effectiveness of multimodal features. This is achieved by exploring different levels of integration of multiple information channels, which, in the case of people close ups videos, correspond to text, audio and video. For a start, the three types of features are use jointly in the classification by merely including the corresponding columns in the same feature matrix. This approach is going to be referred to simply as "features concatenation" (method 3.3.1) and consists in an early fusion of the features. Subsequently, a more sophisticated way of multimodal features combination, based on finding correlation in the way independent cells, belonging to different feature types, are triggered in a neural network to make a prediction. This kind of technique is implemented in the Memory Fusion Network [12]. Let us refer to the latter approach as "features integration" (method 3.3.2).

A more detailed description of the methods that were considered follows below.

#### 3.3.1. BASELINE: TEXT SUMMARIZATION

A trivial text summarization approach is chosen as baseline. In this case, a text summarisation algorithm is chosen for the summarisation of the official transcripts of the speeches from the dataset. From the transcripts summaries it is possible to lead back to a set of chosen segments from the speeches videos. This selected segments are concatenated to form the first type of highlight clips, referred to as "baseline highlight clips".

This approach is purely content based, in the sense that the only feature type upon which the automatic highlights are extracted is textual features from the speeches transcripts, that corresponds to the content. An implementation of the BM25-TextRank algorithm, available in the library gensim<sup>3</sup>, is used for the summarization.

TextRank [131] is a graph-based ranking algorithm: sentences are the vertices in the graphs, the links correspond to the connection between similar sentences, the vertices are characterized by a score that grows with their number of connections. The similarity between two sentences is determined by their content overlap. The algorithm ranks the sentences according to their scores and selects a predetermined percentage of them. The method is fully unsupervised. The variant of TextRank used in this work differs from the original algorithm in the similarity function [132]. More recent Information Retrieval ranking functions, such as BM25

<sup>3</sup><https://radimrehurek.com/gensim/summarization/summariser.html>

and BM25+, resulted in an improvement of 2.92% over the original TextRank. The BM25 similarity between sentences  $S$  with sentence  $R$  is defined as

$$BM25(R, S) = \sum_{i=1}^n IDF(s_i) \cdot \frac{f(s_i, R) \cdot (k_i + 1)}{f(s_i, R) + k_i \cdot (1 - b + b \cdot \frac{|R|}{avgDL})},$$

where  $f(s_i, R)$  is the frequency of the words from  $S$  in  $R$ ,  $k$  and  $b$  are constants,  $avgDL$  is the average length of the sentences.

The key sentences selected by the BM25-TextRank algorithm from each speech transcript are used to retrieve the corresponding video segments, which are then concatenated following the original chronological order to form the video highlights.

### 3.3.2. MULTIMODAL FEATURES CONCATENATION

The first approach of multimodal Machine Learning that is investigated corresponds to combining information from multimodal features through a early fusion. This corresponds to Machine Learning models that are trained with a single feature matrix whose columns are given by the concatenation of different feature types, namely descriptors for text, audio and video information, derivable from the dataset.

The approach of multimodal features concatenation, as well as other technical choices, are inspired by "Harnessing A.I. for Augmenting Creativity: Application to Movie Trailer Creation" [11]. To this extent, the dataset videos are divided into self-standing segments containing complete sentences. For each segment, the same text, audio and video feature are collected and combined, thus obtaining a large feature matrix. The dimensionality of the input matrix is reduced, mainly with PCA, and the resulting features are used to cluster the segments. The names of the clustering algorithm used in the paper are not mentioned, therefore different ones are experimented in this thesis. In addition to this, the video segments are classified also using traditional Machine Learning models, like Random Forests. Ideally, the salient segments, namely the segments that are likely to appear on the news, should be identifiable in the same cluster and be classified as relevant, in the same fashion as the scenes from horror movie trailers analysed in [11].

Out of different algorithms, the best performing one is be chosen. According to the classification made by the latter algorithm, predicted relevant scenes are identified and concatenated in order to form the automatic highlight clips.

### 3.3.3. MULTIMODAL FEATURES INTEGRATION WITH MFNS

The second multimodal approach for video segments classification is based on the neural architecture presented in [12], the Memory Fusion Network, that is described in section 2.4.4. The MFN model allows for a full integration of multimodal features, that is neither a early fusion (features concatenation) or a late fusion (combination of predictions from different modalities). As already seen in the literature review, the MFN consists in three LSTM network (system of LSTMs) that process each feature modality independently. In addition to this, an attention mechanism, the Delta-memory Attention Network (DMAN), that finds correlations between the memories of different LSTM. These correlations are stored in the Multi-view Gated Memory. Therefore, the final prediction is performed combining both the output of the system of LSTM and the combination coefficients calculated by the DMAN. This allows to model both "view-specific interactions" and "cross-view interaction", which are lost in simple feature concatenation.

This more sophisticated approach for multimodal learning is supposed to outperform the feature concatenation described in 3.3.1. The MFN is used to classify the video segments based on their saliency. The segments that are predicted as salient are concatenated to form the "MFN highlight clips".

## 3.4. EVALUATION

The evaluation of the results obtained by the methods deployed is used to demonstrate whether the research questions in 1.2 are true. In fact, if the quality of the automatically generated highlight clips results sufficient, it can be concluded that Machine Learning is able to analyse videos and extract salient information, namely the highlights. In addition, if the results of the evaluations methods show that 3.3.1 and 3.3.2 are superior to the baseline, it is demonstrated that the learning process benefits from the inclusion of features describing different information modalities.

The evaluation of the results is conducted on two levels. First, an automatic evaluation is performed exploiting the saliency annotations available in the dataset. This way, the classification accuracy can be calculated. The classification accuracy represents to which extent the Machine Learning models utilised for the

classification are able to model the relations that characterise the important video segments, according to the saliency labels. The models accuracy can be expressed in terms of precision and recall. The second type of assessment of the results involves human evaluation. In fact, to assess the quality of content of the highlight clips in a high level way, human judgement is necessary. Both types of evaluation are applied to the speeches from the test set, which are never considered during the training phase.

### 3.4.1. AUTOMATIC EVALUATION

The first low-level evaluation is simply conducted by measuring the accuracy of the Machine Learning models, in classifying correctly the video segments. Training the multimodal models in a supervised fashion and calculating the models accuracy is possible because of the presence of the ground-truth binary saliency labels in the dataset. The metrics used to measure the accuracy are precision and recall.

Precision and recall are common metrics, jointly used, especially to assess information retrieval systems and in binary classification problems [133]. In this project, we are testing the effectiveness of Machine Learning to retrieve the relevant video segments from a full video, therefore precision and recall are two suitable evaluation criteria.

Considering the classification results, the true positives (TP) are the video segments labelled as "salient" that actually result classified as salient. The false positives (FP), are non salient video segments that get classified as salient. Similarly, the true negatives (TN) and the false negatives (FN) are, respectively, non salient segments that get classified as non salient, and salient segments that get classified as non salient. Given these definitions, "precision" is calculated as:

$$P = \frac{TP}{TP + FP},$$

that is the fraction of truly salient segments out of all the segments classified as salient. On the other hand, "recall" corresponds to:

$$R = \frac{TP}{TP + FN},$$

that expresses the fraction of all salient segments that is actually classified correctly by the classifier.

Finally, the F1-score (F-score) can be used as a measure of accuracy. It is defined as the harmonic average of precision and the recall, that is:

$$F_1 = 2 \cdot \frac{P \cdot R}{P + R}.$$

### 3.4.2. EVALUATION WITH CROWDSOURCING

Human evaluation is implemented in this thesis through the use of crowdsourcing. Crowdsourcing platforms provide useful tools to implement evaluation tasks and an easy way to recruit workers at a large scale. The additional human evaluation allows for the assessment of the highlight clips produced in the thesis in terms of their degree of informativeness, ability to spark interest in the viewers, fluency and coherence in the selected video shots. This level of information is not possible to obtain through automatic evaluations. The crowdsourcing tasks implemented to this end are divided into two crowdsourcing jobs, consisting of surveys with closed-ended questions and Likert questions, which are published on the platform Figure Eight [134] and address English speaking workers. The final design of the surveys was achieved after conducting a pilot evaluation round, that provided preliminary results and workers feedback about the clearness of the questions. The definition of the complete crowdsourcing process is explained in details in chapter 6.

If from the results of crowdsourcing it is possible to extract a fair and complete evaluation of the methods for highlights extraction, then **RQ3** is confirmed to be true. Finally, this high level evaluation allows to measure more precisely the effectiveness of Machine Learning for highlights extraction and two compare the results obtained with the different approaches.





# 4

## NOVEL DATASET FOR HIGHLIGHTS EXTRACTION FROM POLITICAL SPEECHES

### 4.1. MOTIVATION

The goal of the Avengers project is to create a system capable of creating short highlight clips, that represent the most salient content, extracted from the input to the system, as shows by the project pipeline in Figure 1.1. The input to the system is multimodal, and consists of a combination of text, audio or video files, concerning the life of a person, a conference, or other types of events. The purpose of this highlight clip creation task, shared by many other works in the Machine Learning research area about automatic video analysis and summarisation, is to extract relevant information from unstructured data, creating a browsable and, possibly, engaging short version of it that can be shown to an audience. Ultimately, the Avengers project should result in a product that can be used by users to create personalized video clips.

A possible way to reach to goal of the Avengers project is through the replication of the work described in the paper "Harnessing AI for augmenting creativity: Application to movie trailer creation." [11], discussed in §2.1.2. In this case, the problem would be shifted from identifying scenes that are good candidates for a horror movie trailer to identifying the parts of a video that should included in a highlight clip. However, to be able to reproduce the methodology that the authors introduced, there are requirements to satisfy. Firstly, the access to a large amount of testing and training data is needed. The dataset must be a collection of video pairs (long video - shorter version), like in the case of movie - movie trailer. This is easy to find when dealing with movies, but the same type of data might not exist for other kinds of videos. Secondly, the videos should have an underlying structure, or some recurrent characteristics, just like tropes, that can be captured by a statistical model and used to identify the relevant parts in them.

This thesis is focused on the part of the Avengers project that concerns the analysis of videos where one person speaks in front of a camera, like the movie reviews contained in the MOSI and MOSEI datasets [37, 38]. Video analysis for this kind of videos often includes sentiment classification, due to the possibility to extract cues like facial expressions, gestures, changes in the intonations. For this reason, sentiment analysis has been considered for the task of highlights extraction from videos where a person speaks to the camera, under the hypothesis that relevant scenes contain some emotional patterns that can be detected. However, finding a dataset containing videos like the ones formerly described is not easy and finding a collection of videos where one person speaks to the camera with annotations regarding ground-truth summaries of relevant video segments is harder. Therefore, in this case the methodology from [11] could not be applied. Some other methods have been considered, but a direct deployment of Deep Learning algorithms is not applicable due to the lack of data, and transfer learning from another domain is likely to produce poor results. In addition, resorting to crowdsourcing to create this sort of dataset requires excessive time and monetary resources, not to mention a lot of effort from the crowdworkers in watching hours of videos and summarizing them manually, a task that not everyone might be willing to execute professionally.

The need of data led to a search that found an answer in political speeches and news clips. Political speeches are always professionally recorded and shared online on different platforms, like YouTube, news websites or political party websites. Moreover, they usually have an official transcription. As an example, the website "American Rhetoric - An Online Speech Bank" offers a collection of all the speeches from Barack

Obama and other politicians, with the official transcription<sup>1</sup>. In the politics section in the news, very often highlights from the speeches of politicians are shown. In addition, news channels often upload on their websites videos that contain the highlights of important political speeches. Putting aside the political bias, it can be safely assumed that these highlight clips, which usually last for a few minutes, represent the most important moments in a speech that is likely to last more than one hour. In addition, in most cases speeches follow a basic structure, that can help identifying their focal points. The fact that politicians employ body languages techniques to enhance the credibility or attention grabbing strategies when they want to say something important might be useful as well. To sum up, political speeches are structured, widely available, open source, with good quality filming and transcripts. The availability and the supposedly structured nature of political speeches make them good candidates for a dataset used to train a model for highlights extraction. This dataset could be deployed for interesting computations, experimenting with different Machine Learning architectures and features. With this respect, the thesis objective is the detection of saliency in political speeches and its application in automatic highlight clips creation.

To achieve the thesis objectives, a novel dataset was created, containing video speeches from American politicians and the respective highlight clips broadcast on American news channels, from which it is possible to extract "saliency labels" which describe the importance of a video segment in the context of the full speech.

## 4.2. CREATION OF THE DATASET

The target dataset is collected starting from a list of political speeches by different politicians. Three essential elements were identified and required for each political speech:

1. full length speech video,
2. highlight video made by a reliable news channel,
3. official transcription of the speech.

In order to collect all the three elements, first reliable and impartial sources of information about political speeches were researched. The two main sources that were found are the freely accessible websites American Rhetoric<sup>2</sup> and Factbase<sup>3</sup>. The former website is meant to be a support for practicing rhetorical skills and contains a collection of speeches of political, social, movie and religious nature, that are usually proposed with a film or audio record and the transcription. The latter website is a platform that conducts analytics of political matters in different ways, including sentiment in congresses or tweets, transcription, voice and video analysis of speeches and prediction of the personality traits of some political leaders. Factbase shares all Donald Trump's transcribed speech segments, with videos and sentiment scores and is always up to date with the latest speeches.

From American Rhetoric, speeches from Barack Obama, Hillary Clinton and George W. Bush were collected, while speeches from Donald Trump were collected from Factbase, for a total of 99 political speeches. The choice of this number is motivated by the fact that in [11], 100 trailers were used to construct the statistical model. The list of 99 speeches was composed manually, checking for every speech the quality of the full speech video and the presence of a good highlight video clip and an official transcription. The chosen speeches are characterized by the fact that there is only one politician speaking in the scene, except for some rare cases where the main speaker is introduced by another person, or the speech is part of an event where more guest speakers are involved. Videos of speeches that did not meet these requirements were discarded.

Each speech is unequivocally identified by a speech title, derived from the speeches sources together with the official transcript. The speech title and speech date were used to retrieve an official video of the full speech from YouTube and an official highlight clip, shared by a news channel or a political channel. A list of all 99 speeches, with speech titles and links to the full speech video, highlight clip and transcription can be found in the Appendix A.1. The full videos are focused on the main speaker for almost their whole duration, interrupted by the inclusion of some brief shots of the audience. The scenes are quite static, with a plain or uniform background that makes possible to identify and isolate the speaker's face. The audio and video quality are good enough to have clear images and sounds. On the other hand, the highlight video clips often show news presenters and commentators that alternate to fragments from the political speech in question. In addition, in most cases the shots from the highlight video clips do not strictly match the ones from the

<sup>1</sup><https://www.americanrhetoric.com/barackobamaspeeches.htm>

<sup>2</sup>Available at <https://www.americanrhetoric.com>

<sup>3</sup>Available at <https://factba.se>

full speech videos, as they might be recorded by different cameramen, or they might show archive images or videos with the speech in the background.

#### 4.2.1. LABELLING PROCESS

Exploiting the highlight clips, saliency labels for each speech segment were extracted. Here, the term *saliency* is used to define moments in the video of a political speech that are considered important, because they also appear on at least highlight video clip broadcast by a reliable news channel. The process adopted to derive the video segments and the saliency labels is described as follows.

First, both the full-length videos and highlight clips of all the speeches were transcribed using Google Cloud Video Intelligence<sup>4</sup>. This API returns a JSON<sup>5</sup> file containing a list of punctuated text segments with timestamps for each occurring word. The output transcript is segmented according to the pauses in the speech. In order to evaluate the accuracy of the transcription from Google Cloud Video Intelligence, the ROUGE score [135] between the official transcript and the automatic transcript of the full speeches was calculated. The results are shown in Figure 4.1, which shows precision, recall and f-score according to the metrics ROUGE-1, ROUGE-2 and ROUGE-l, which consider, respectively, 1-gram, 2-grams and the longest common subsequence. Possible lower scores are due to small spelling mistakes, often due to the presence of proper nouns. However, the resulting ROUGE scores are high enough to consider the Google Cloud Intelligence Transcription reliable.

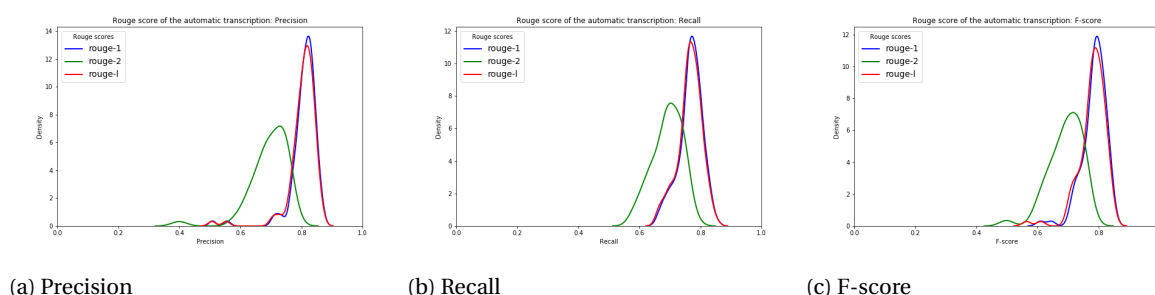


Figure 4.1: ROUGE scores obtained by the transcription with Google Cloud Video Intelligence compared to the official speech transcripts

The videos of the full speeches were divided into segments according to the segmentation in the transcription, exploiting the timestamps. This way it is guaranteed that the video segments contain sentences with complete meaning and no truncated words. The saliency labels were derived from the overlaps between full speech videos and highlight clips, under the hypotheses that the "salient content", namely the most important part in a political speech, are the parts broadcast to be commented and discussed in the news. In order to find the overlaps, keeping in mind that the images are likely not to match, so the visual content cannot be exploited, the automatic transcripts were used. The difficulty of this task is due to the fact that Google Cloud sometimes fails to transcribe some words correctly, therefore may transcribe the same words pronounced in the full-length video and highlight clip in two different ways. For example, the word *anyway* could be transcribed as "anyway" or "any way", or the word *there* could be erroneously transcribes as "their" or "they're", but the presence of these kinds of mistakes are hard to predict, localise and correct automatically. In addition, the automatic segmentation of the full length speeches videos and the corresponding highlight clips might be not exactly the same, so 100% matching between the transcripts of the segments of these the two video types is impossible to obtain.

Nevertheless, a function was implemented to check which parts of the full speeches are present in the correspondent highlight video. The function measures the overlapping between two sentences, keeping into account eventual transcription mistakes: if the overlapping percentage is above a certain threshold, the function detects a match. This way, it was possible to assign to all the segments of the full-length videos binary "saliency labels", namely 0 for segments that do not appear in the video highlights, and 1 for the "important" segments that are reported in the video highlights. The labels for each speech were stored in NumPy arrays<sup>6</sup>. Ultimately, in order to observe the quality of the resulting labels, automatic video highlights were composed for each speech, concatenating the video segments labelled 1. These automatic highlight clips, included in

<sup>4</sup>Available at <https://cloud.google.com/video-intelligence/>

<sup>5</sup>JSON: <https://www.json.org>

<sup>6</sup>NumPy documentation: <https://www.numpy.org>

the dataset, are considered the "ground truth" highlights. The replication of these ground truth highlights in an automated fashion, using Machine Learning or Deep Learning models, is the objective of the thesis experiments.

The whole process, from the videos downloading, the transcription videos transcription, to the overlapping detection and the creation of the highlight clips was all automated and managed through a code written in Python.

#### 4.2.2. MULTIMODAL FEATURES EXTRACTION

Following the methodology of [38] and [12], multimodal features were extracted from the video segments. These include word embeddings from the speech transcripts, audio features from the speech signal processing and descriptors for the facial expressions and head pose from the visual content. The process of feature extraction was automated and deployed different software.

Pre-trained GloVe vectors [106] from two different corpora were used for the text representation: "Wikipedia 2014 + Gigaword 5", containing 6 billion tokens, and "Common Crawl", containing 840 billion tokens of web data<sup>7</sup>. In this case, the transcript computed by Google Cloud Video Intelligence was used instead of the official transcript: this is motivated by the fact that small changes between the automatic transcription and the official transcription, e.g. the spelling of the word "anyway" separating "any" and "way", or the allocation of the word "as" for "is" and vice-versa, make it difficult to segment the official transcript in the same way as the automatic transcription, which is mirrored in the video segmentation. However, the two versions of the transcripts are sufficiently similar to have good word representations, as proven by the calculation of the ROUGE scores [135]. As expected, the use second pre-trained vector resulted in a larger number of represented words: 595 words were not using "Wikipedia 2014 + Gigaword 5" and only 428 with "Common Crawl". For this reason, the latter pre-trained vectors is employed for the extraction of text features. In the majority of cases, the missing words correspond to proper nouns or numerical data. The list of missing words using "Common Crawl" is reported in the Appendix A.2.

As for what concerns the audio features, an extensive speech analysis was performed with COVAREP [108]. The features extracted with the open-source code include prosodic features, voice quality features and spectral features, sampled at 100 Hz.

Finally, the software OpenFace 2.0 [136] was used to extract facial landmark detection, head pose estimation, facial action unit recognition, and eye-gaze estimation. It is noticeable that the speakers' faces are not framed for the whole duration of the speeches, as sometimes the cameramen turn to the audience. However, when the algorithm is not able to detect a distinct face, automatically sets the output to 0. In addition, it is important to underline the fact that OpenFace is extremely performing even when the video quality is not excellent or when more people other than the main subject are visible in the background. An example of faces detected and isolated by OpenFace is shown in Figure 4.2

Once all the three types of features were computed, it was necessary to align the different features vectors in order to make sure the three input modalities are synchronized in time. In fact, text, audio and visual features are sampled at different rates: one GloVe embedding is extracted for every pronounced word, COVAREP features are sampled with a frequency of 100 Hz and OpenFace features are computed for every video frame. The videos do not share the common frame rate. Since the sampling frequency of the word vector is the lowest, alignment was performed by averaging the audio and visual features falling between the pronunciation of one word. This was possible as the timestamps for all the units of the three feature types was known: the transcripts word timestamps are output during the speech transcription, the OpenFace features contain the timestamps and the sampling rate for COVAREP was manually set. As one segment is described by a collection of multimodal feature vectors for each time unit, and the labelling process is performed segment-wise, to each feature vector the same label as its corresponding segment was assigned.

<sup>7</sup>Available at <https://nlp.stanford.edu/projects/glove/>



Figure 4.2: Some results output by OpenFace [136] during the feature extraction.

### 4.3. CONTENT OF THE DATASET

This section provides more details about the Political Speeches Dataset content and structure.

The final dataset contains data for 99 speeches:

- 51 videos for Barack Obama,
- 38 videos for Donald Trump,
- 4 videos for George W. Bush,
- 3 videos for Hillary Clinton,
- 1 video for James B. Comey,
- 1 video for Nikki Haley.

The data collected for each speech comprises:

- the official transcript of the speech,
- the video of the full speech,
- one highlight clip created for the speech by one news channel,
- text segments of the automatic transcription extracted with Google Cloud Video Intelligence,
- video segments of the full video that reflect the segmentation of the automatic transcription (based on the pauses in the speech),
- a vector of binary "saliency labels", each corresponding to one segment (0 for "non salient" segments and 1 for "salient" segments),

- a set of 300-dimensional GloVe embeddings extracted from the automatic text segments,
- a set of 74-dimensional audio features extracted with COVAREP,
- a set of 714-dimensional visual features extracted with OpenFace.

The processing on the dataset results in a total of 15624 labelled video segments with respective multimodal features. Each speech has an average of 158 segments, with standard deviation 102. Figures 4.3, 4.4 and 4.5 report some information about the duration of full-length videos, highlight videos and video segments.

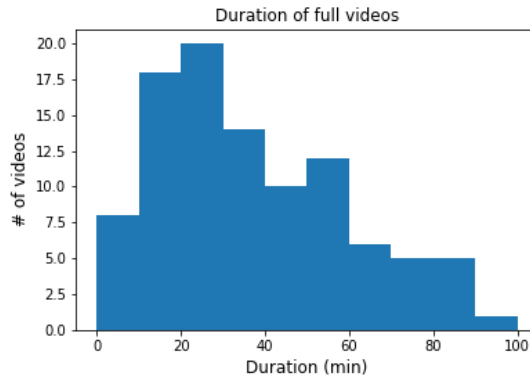


Figure 4.3: Duration of the full-length videos. Mean duration: 37.70 min, std: 22.33 min

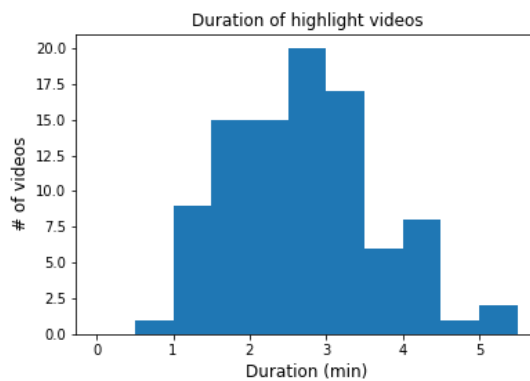


Figure 4.4: Duration of the highlight clips. Mean duration: 2.99 min std: 1.59 min

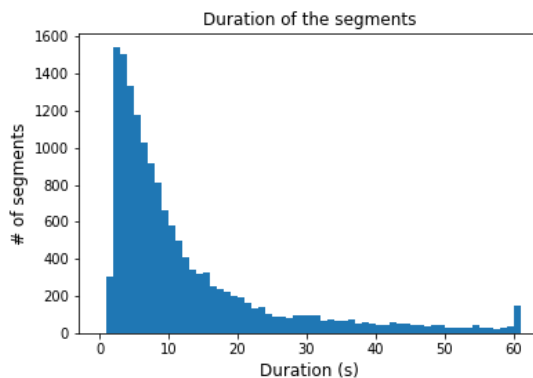


Figure 4.5: Duration of the video segments. Mean duration: 12.81 s, std: 13.28 s. Minimum duration: 1.02 s, maximum duration: 61.03 s

# 5

## EXPERIMENTS

### 5.1. INTRODUCTION

This chapter presents the computational experiments, conducted to investigate the effectiveness of the methods explained in Chapter 3, and the respective classification results. The experiments involve three main Machine Learning approaches:

1. unimodal classification: Baseline method (§5.3);
2. multimodal features concatenation: Scene clustering/classification with traditional Machine Learning (§5.4);
3. multimodal features integration: Saliency prediction with Deep Learning using MFNs [12] (§5.5).

The first section, §5.2, is dedicated to the explanation of the preprocessing of the data, which is fundamental to obtain a non redundant feature matrix, which contains compact and discriminative features.

### 5.2. PREPROCESSING OF THE DATA

Data preprocessing is a preliminary operation to the classification with Machine Learning and Deep Learning algorithms. Preprocessing typically involves data cleansing, normalisation and dimensionality reduction. These steps are necessary to eliminate redundancy in the feature matrix, which results in an improve of the algorithms accuracy, as well as of the computational speed.

Before conducting any analysis of the data, the dataset was split into training and test set. The splitting was done in a way to preserve the components of one speech in the same set. 70 random speeches were included in the training set and 29 in the test set. This resulted in 320045 data points in the training set and 123609 in the test set. Each data point corresponds to a multimodal feature vector sampled at the time unit of one pronounced word.

From Figure 5.1, which contains the partial visual representation of the feature matrix, it can be noticed that the three feature types, namely text, audio and video features, differ in terms of number of components and scale. This observation is confirmed by the visualisations of the mean and standard deviation of the different feature types (Figures 5.2a to 5.2f), that show that both the inter-variance and the intra-variance assume a wide range of values.

For this reason, it is necessary to normalise or standardise the data. Standardisation consists in centering the data in zero and rescaling the data in order to obtain unitary standard deviation, following the formula  $z = \frac{x_i - \mu}{\sigma}$ , where  $x_i$  represents an instance of feature  $i$ ,  $\mu$  is the mean value of the distribution of feature  $i$  and  $\sigma$  its standard deviation. The reason why standardisation was applied to the feature matrix is that many Machine Learning algorithms, including Principal Component Analysis, assume zero centered distributions of the data. Also, without bringing the data to a common scale, some features ranging over a larger scale would automatically drown out the other features, even the ones that are potentially more discriminative.

Figure 5.3 shows the 100 first rows of the full feature matrix after applying standardisation. On first glance, it can be inferred that the matrix contains redundant features, as for the features columns approximately from 500 to 650 maintain a constant colour.

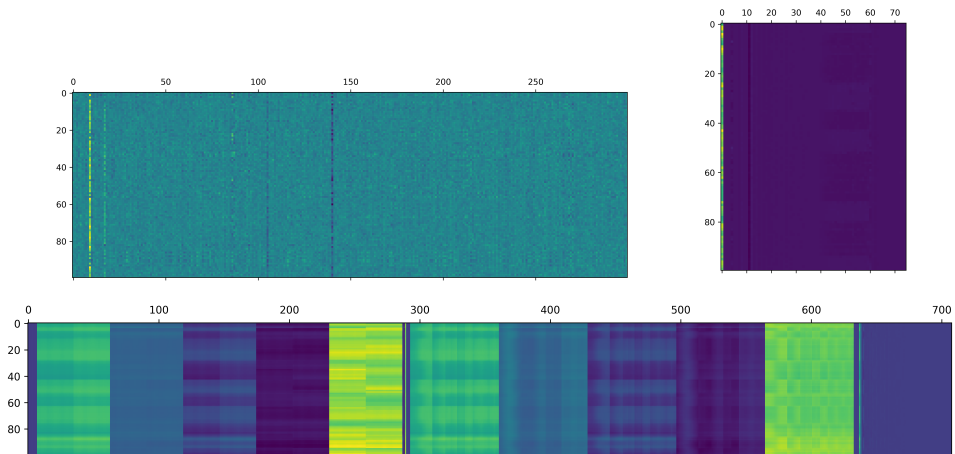
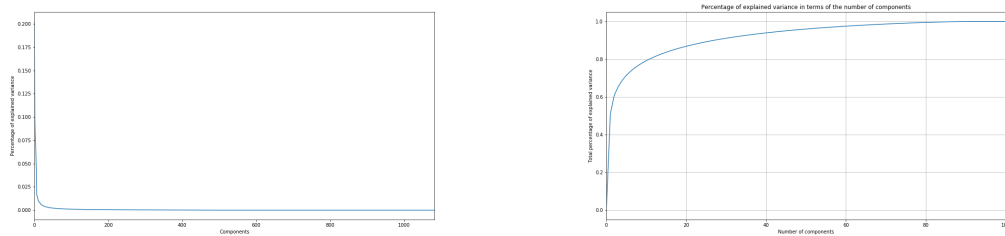


Figure 5.1: First 100 rows of the feature matrix for text, audio and visual features

### 5.2.1. DATA DIMENSIONALITY REDUCTION

Principal Component Analysis (PCA) is a dimensionality reduction technique that allows to map a feature matrix into a lower-dimensional matrix that maintains the essential data patterns or the original matrix [137].

PCA was applied to the standardised training set, both on the full feature matrix and on the partial matrices for each of the three feature types. Figure 5.4a shows the percentage of variance "explained" by each of the 1083 components, sorted by their "explainability ratio". It is clear how the components that follow the first one, which explains about 20% of the total variance, rapidly become non informative. The next plot, Figure 5.4b, shows how the total percentage of explained variance varies as more components are added. It can be observed that 90% of the total variance is achievable with only 27% of the number of components (approximately 292 features out of 1083), 95% with 45% of the number of components and 98% with 65% of the number of components.



(a) Percentage of explained variance for each of the 1083 features, sorted by their "explainability ratio"

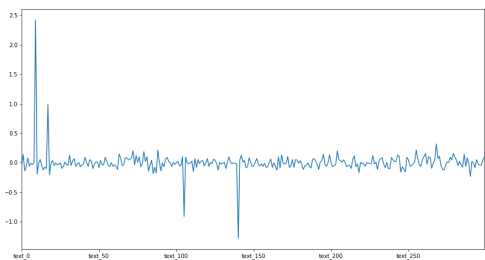
(b) Total percentage of explained variance as the number of components increases

The same analysis was conducted on each feature type individually. Figures 5.5, 5.6, 5.7 and 5.8 show the different behaviour of text, audio and video features. If in the case of text features, 95% of the components is necessary to explain 95% of variance, the same is obtainable with only 9% of components for audio features and 6% of components for video features. This underlines the heterogeneity of the multimodal feature space and the fact that many information extracted with the COVAREP<sup>1</sup> and OpenFace<sup>2</sup> softwares are not important for the classification task of the thesis.

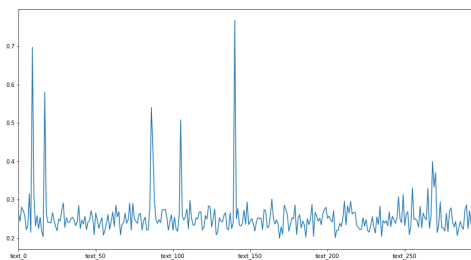
<sup>1</sup><https://covarep.github.io/covarep/>

<sup>2</sup><https://cmusatyalab.github.io/openface/>

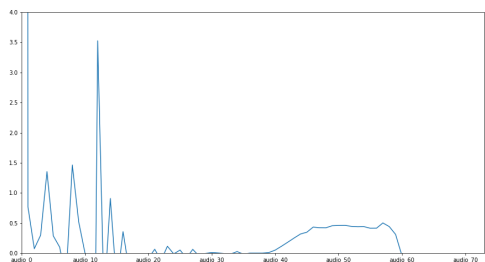




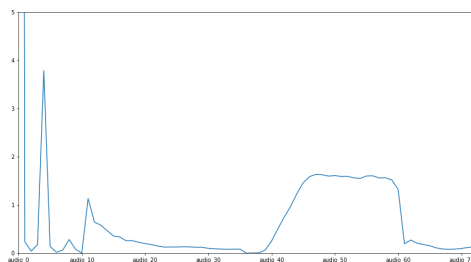
(a) Text features: mean



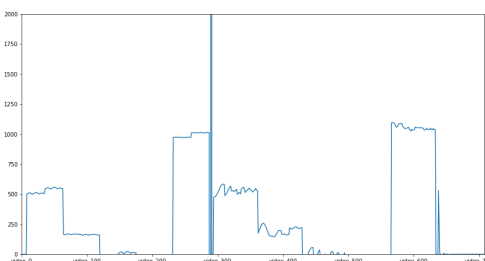
(b) Text features: standard deviation



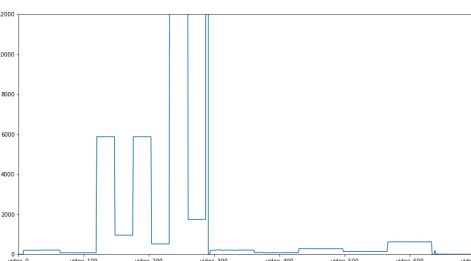
(c) Audio features: mean



(d) Audio features: standard deviation



(e) Video features: mean



(f) Video features: standard deviation



Figure 5.3: First 100 rows of the full standardized feature matrix

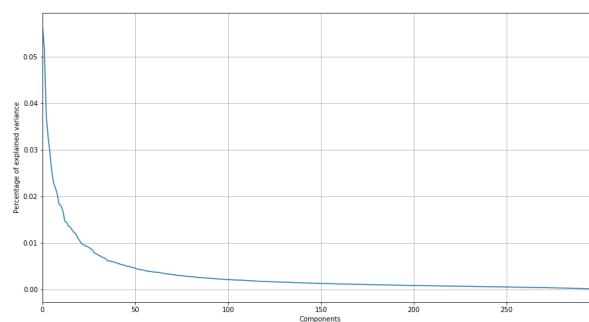


Figure 5.5: Percentage of explained variance for text features

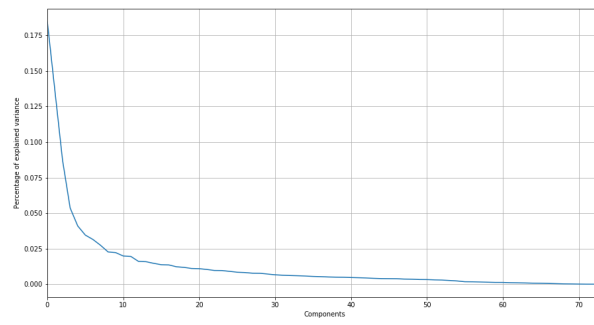


Figure 5.6: Percentage of explained variance for audio features

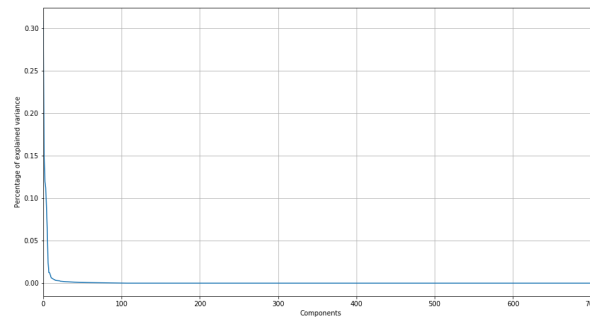


Figure 5.7: Percentage of explained variance for video features

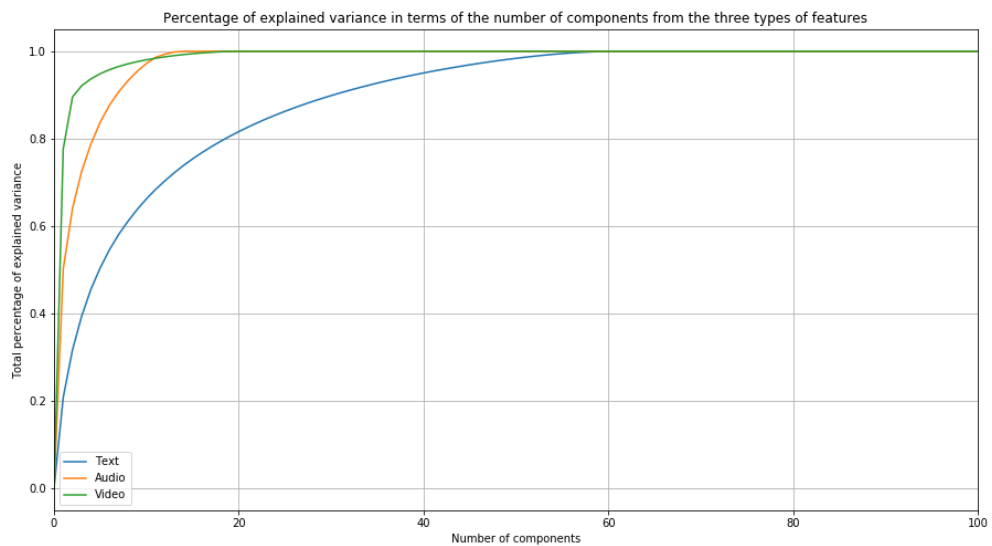
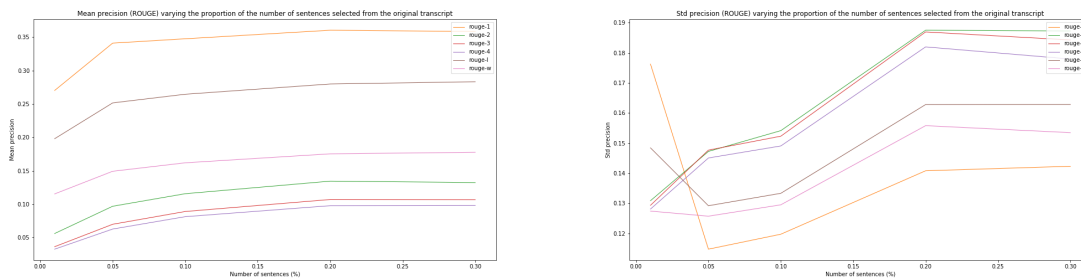


Figure 5.8: Total percentage of explained variance for each component type

### 5.3. BASELINE METHOD

The first method that is analysed is the baseline, namely the unimodal approach based on the text summarisation of the official speech transcripts, using the TextRank [131] algorithm. The algorithm was applied to the official transcripts of all the speeches from the Political Speeches Dataset, in order to produce extractive summaries for each. To measure of the performance of the baseline, the produced transcript summaries were compared to the summaries of the ground truth highlight clip, and the ROUGE score was calculated. This was done for multiple values of the parameter describing the summary length.

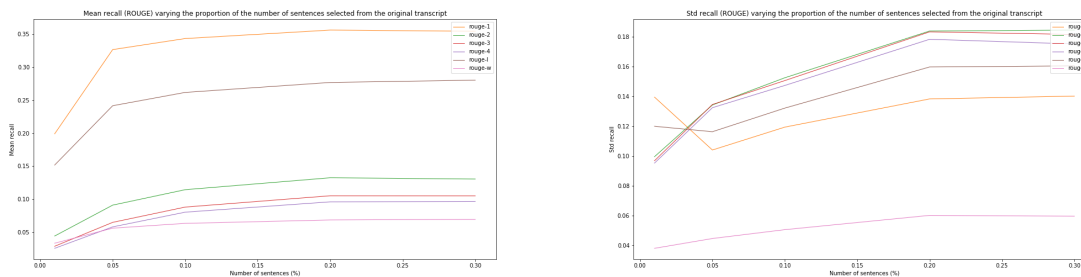
Figures 5.9, 5.10, 5.11 show the mean and the standard deviation of precision, recall and F-score, calculated with different ROUGE metrics<sup>3</sup> for all the speeches in the dataset. ROUGE- $n$  ( $n \in \{1, 4\}$ ) measures the overlap of  $n$ -grams, ROUGE- $l$  is based on the longest common subsequence (LCS) and ROUGE- $w$  on the weighted LCS. As can be seen in the graphs, the best mean scores obtained for precision, recall and F-score according to ROUGE-1 is 0.36, with standard deviation 0.14. The low quality of this result, can be motivated by the fact that the ground truth highlight clips and the transcripts summarisation result in two presumably different approaches for highlights extraction. The criteria used by the professional filmmakers that work for the news channels, might not be based on text summarisation. The political bias might also affect these differences. In addition to this, the small transcription mistakes committed by Google Cloud Video Intelligence negatively affect the ROUGE scores. Eventual limitations of the BM25-TextRank algorithm itself might also be a cause for low ROUGE scores. Nevertheless, from the output transcript summaries it was easy to identify the corresponding video speech segments. These were concatenated to produce the so called "baseline highlight clip". This simple unsupervised method for highlights extraction and highlight clips creation is considered the baseline, to be outperformed by the other multimodal Machine Learning approaches. Because of the reasons previously discussed, it would not be fair to compare the multimodal methods with the baseline in terms of ROUGE scores. In fact, the superiority of the other types of highlight clips needs to be tested through human evaluation.



(a) Mean precision

(b) Std precision

Figure 5.9: ROUGE scores of the baseline method: Precision

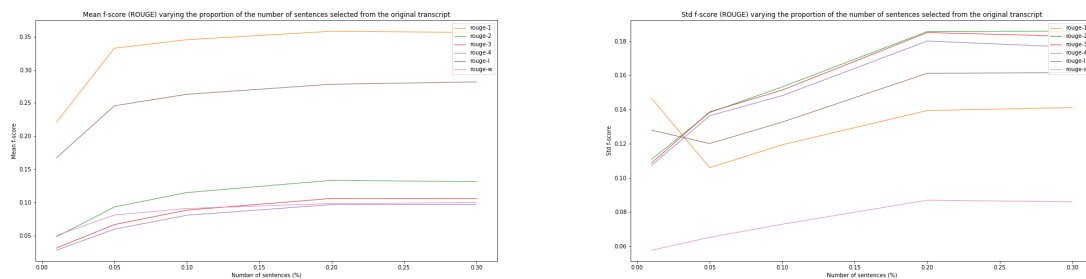


(a) Mean recall

(b) Std recall

Figure 5.10: ROUGE scores of the baseline method: Recall

<sup>3</sup>Python implementation available at <https://pypi.org/project/py-rouge/>



(a) Mean F-score

(b) Std F-score

Figure 5.11: ROUGE scores of the baseline method: F-scores

## 5.4. MACHINE LEARNING MODELS WITH MULTIMODAL FEATURES CONCATENATION

Following the methodology from [11], the first attempt of classification was conducted applying two main unsupervised clustering algorithms: Mini-Batch K-Means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). As the data analysis showed that a large percentage of feature components can be discarded as they are barely informative, the classification was conducted after reducing the dimensionality with PCA. Since this approach did not lead to meaningful results, another classification method based of the Random Forest algorithm was attempted, resulting in accuracy of 77.42%.

For all the classification methods and the preprocessing, some of the models implemented in the library scikit-learn<sup>4</sup> were used.

### 5.4.1. MINI-BATCH K-MEANS CLUSTERING

K-Means clustering consists in the identification of  $n$  disjoint clusters, where  $n$  is a fixed given number, described by the mean of the samples in the cluster (the *centroid*,  $\mu_j$ ), such that the inertia of clue cluster is minimized. That is:

$$\sum_{i=0}^n \min_{\mu_j \in C} \|x_i - \mu_j\|^2$$

Mini-Batch K-Means clustering is a variant of the K-Means clustering algorithm used to reduce the computational time by performing the clustering with iterations of randomly sampled mini-batches.

The algorithm was initially applied to the unbalanced training set varying the number of components selected by PCA (from 1% to 50% for each feature type) and with  $n = 2$ . As expected, this naive approach results in an average accuracy that ranges between 32% and 68%, where the highest score is obtained by assigned the majority of the points to the cluster labelled 0, since almost 93% of the data belong to this class. The balances accuracy (average of recall obtained on each class) is not larger that 54%. This poor results are explainable with the fact that the two classes are highly inseparable and overlapping, therefore is it not possible to divide them in only two single clusters.

Given this, the same algorithm was applied iterating over different number of clusters  $n$  (2, 3, 5, 10, 20, 30, 40, 50). This time a sample of 40000 data points was randomly extracted from the training set, in order to have 50% points from each class. In addition, a constant number of feature components, 197, was selected such to retain at least 90% of the total variance from text, audio and video features. Ideally, the clustering should result in  $n$  clusters that present a significant prevalence of elements labelled 0 or labelled 1. However, for all the values of  $n$  such detection was impossible: in fact, almost all the identified clusters contain approximately the same amount of data points from class 0 and class 1, as Figure 5.12 shows.

The same classification was attempted considering only one feature type at the time. PCA was applied in order to include an amount of components that allow the representation of at least 95% of the total variance, that is 200 components for the text features, 44 components for audio feature and 26 components for video features (see Figure 5.8). However, as can be seen in Figures 5.13, 5.14 and 5.15, no relevant results were obtained. Again, the reason for this is probably that it is impossible to divide the data into clusters as the classes are too overlapping.

<sup>4</sup><https://scikit-learn.org/stable/>

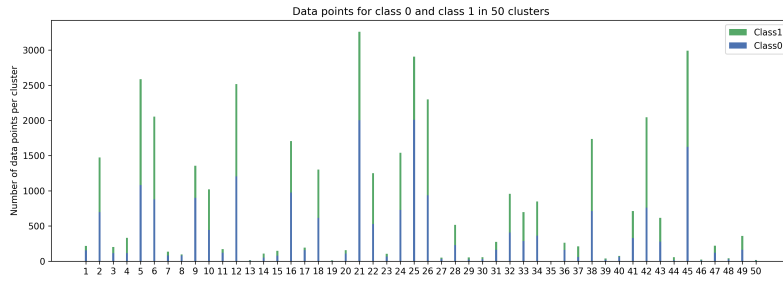


Figure 5.12: Mini-Batch K-Means with  $n=50$ . Each column corresponds to the elements in a cluster and shows the percentage of data points from class 0 and class 1

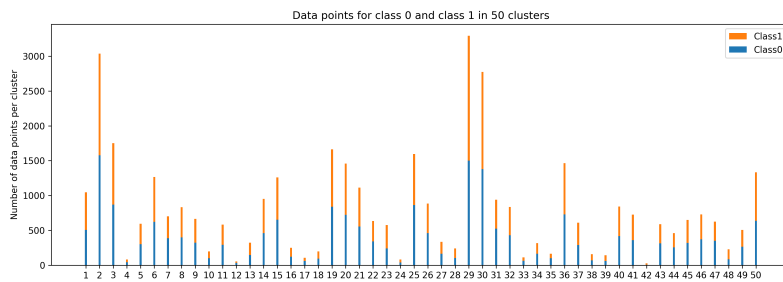


Figure 5.13: Mini-Batch K-Means with  $n=50$ , text features

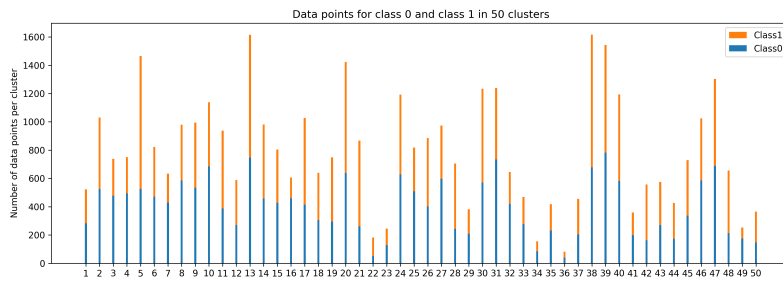


Figure 5.14: Mini-Batch K-Means with  $n=50$ , audio features

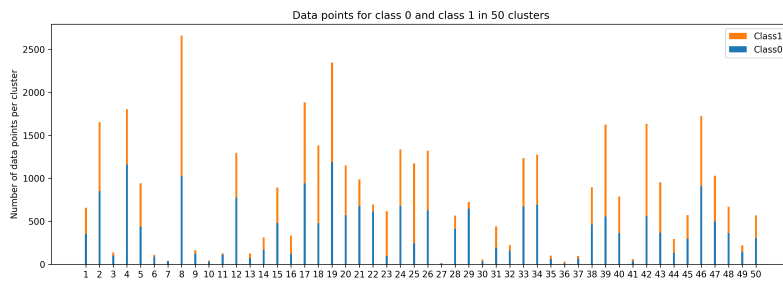


Figure 5.15: Mini-Batch K-Means with  $n=50$ , video features

### 5.4.2. DENSITY-BASED SPATIAL CLUSTERING OF APPLICATIONS WITH NOISE

Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a clustering algorithm that overcomes the need of K-Means of having a predetermined number of cluster, and is also resistant to outliers, that get classified as noise. The algorithm requires two main parameters: the distance threshold  $\epsilon$  and the minimum number of data points *minPoints*. First, one initial data point is selected, and its neighbours are checked. If there are at least *minPoints* neighbours that fall within distance  $\epsilon$  from the initial point, these data points are all assigned to one cluster, otherwise the initial data point is classified as noise. The algorithm continues assigning the data points to existing clusters, if the distance from their components is less than  $\epsilon$ , otherwise new clusters are created. The algorithm iterates over all the points in the training set until they are all marked as visited.

DBSCAN was trained only on data points from class 1, which is the target, namely the class the corresponds to the speeches highlights. Ideally, after clustering, validation data points from class 1 should be assigned to the predicted cluster(s) and data points from class 0 should be classified as noise. 17802 of class 1 were sampled from the training set, in order to guarantee 5000 points in the validation set from both classes. The same 197 feature components as for K-Means were used. The training was performed with different values for  $\epsilon$  (0.1, 1, 10, 20, 30) and *minSamples* (2, 3, 5, 10, 20, 30). Intuitively, with unitary *minSamples* all the points are assigned to a cluster of one element and the number of clusters is maximum. By increasing *minSamples*, if  $\epsilon$  is large enough the clusters are agglomerated and the number of clusters decreases. However, if *minSamples* becomes too large, it might not be possible to form any cluster and the number of clusters tends to zero. This behaviour is showed in Figure 5.16.

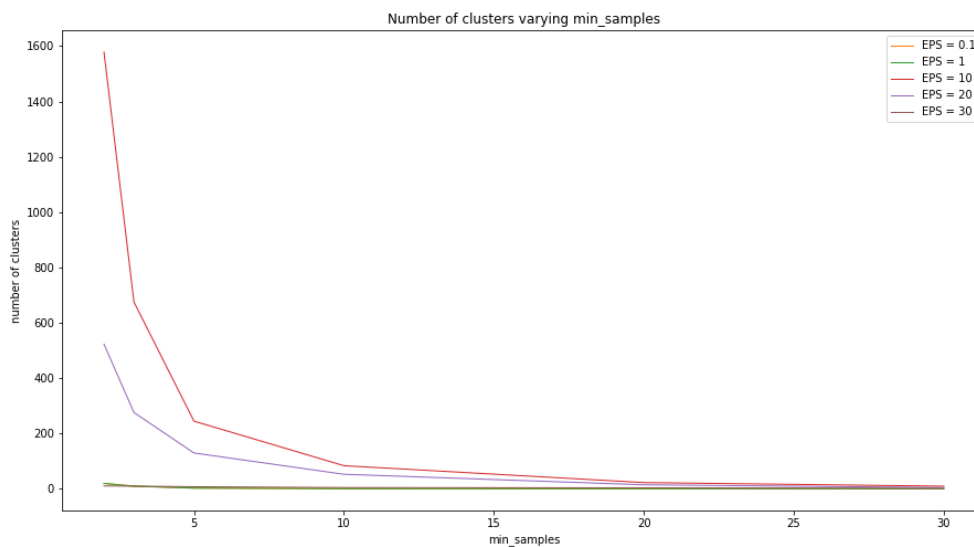


Figure 5.16: Number of detected clusters with different  $\epsilon$  and *minSamples*

The classification was tested on the validation set. As scikit-learn does not provide a "predict" method for DBSCAN, a function that labels the points either as "classified" (to class 1) or "noise" was implemented. The function calculates the Euclidean distance between the candidate point to classify and the clusters cores, and assigns the candidate point to the same group as the first core point that is closer than  $\epsilon$ . Figures 5.17, 5.18, 5.19 and 5.20 show some of the results. The three plots correspond to three values of  $\epsilon$ , namely 1, 10, 20 and 30. Results with other values are omitted as they do not differ from the case  $\epsilon = 10$ . Each bar corresponds to one value of *minSamples*. The left part corresponds to class 0 and the right part to class 1. The stacked bars show the percentage of data from the validation set that is classified as belonging to one cluster (therefore assigned to class 1) or classified as noise, for the two classes. As can be seen, with small values of the distance threshold  $\epsilon$ , all the points get classified as noise (blue and green). By increasing  $\epsilon$ , the points start getting assigned to clusters. However, if *minSamples* is too large with the predetermined  $\epsilon$ , there are not enough data points to form a cluster and these points get classified as noise. If  $\epsilon$  is large enough, all the points get assigned to some cluster. Unfortunately, the classification produces similar results for class 0 and class 1. In

fact, data points belonging to these classes are classified as class 1 or noise with similar percentages in all cases. This confirms the impossibility to cluster the data due to the classes overlapping.

The same approach was repeated only considering one feature type at a time, after reducing the dimensionality in order to retain a number of principal components that explain at least 95% of the total variance (200 for text, 44 for audio and 26 for video). Similarly as in the above case, no significant results were obtained (see Figures 5.23).

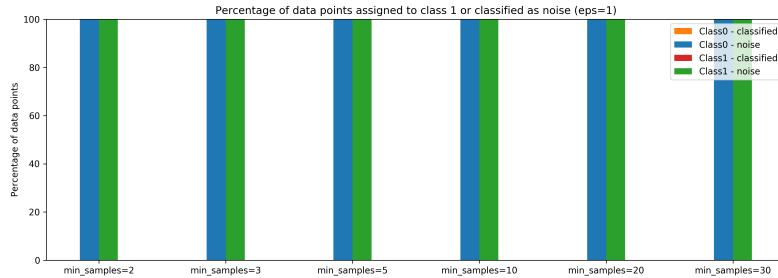


Figure 5.17: Results with DBSCAN,  $\epsilon = 1$ , different *minSamples*

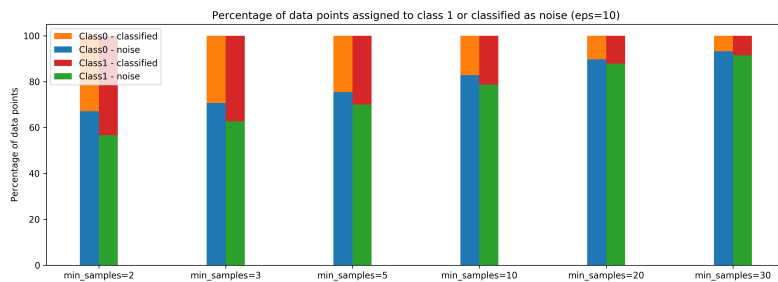
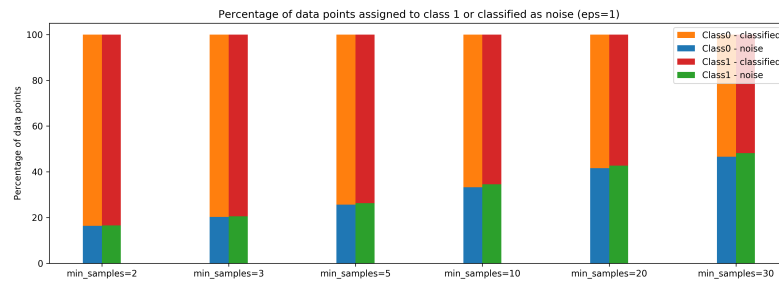
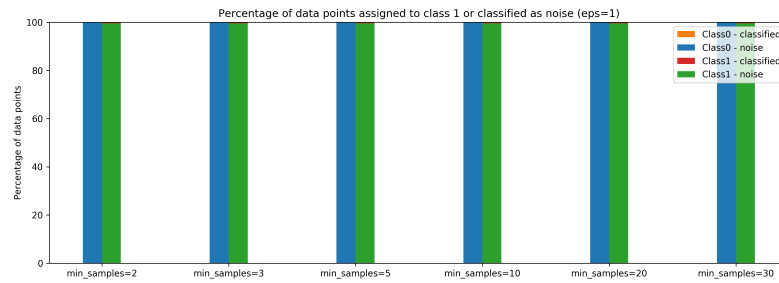


Figure 5.18: Results with DBSCAN,  $\epsilon = 10$ , different *minSamples*



Figure 5.19: Results with DBSCAN,  $\epsilon = 20$ , different *minSamples*

Figure 5.20: Results with DBSCAN,  $\epsilon = 30$ , different *minSamples*Figure 5.21: Results of DBSCAN with text features ( $\epsilon = 1$ , different *minSamples*)Figure 5.22: Results of DBSCAN with audio features ( $\epsilon = 10$ , different *minSamples*)Figure 5.23: Results of DBSCAN with video features ( $\epsilon = 10$ , different *minSamples*)



### 5.4.3. RANDOM FOREST

From the results obtained with K-Means and DBSCAN it appears that PCA followed by unsupervised clustering is not the right track to achieve relevant results. For this reason, another approach was attempted.

First, the standardized training set was divided into class 0 and class 1. The mean and standard deviation were calculated for each feature of the two subsets. Then, the difference between the mean of the features from class 0 and class 1 was derived (Figure 5.24). The features were sorted in descending order, according to the difference in the means (Figure 5.25). The assumption is that in the dimensions where the difference of the means is maximum the two classes are separable, even though the plot of the standard deviations (Figure 5.26) suggests there might still be some overlapping. Given this, the first  $n$  features were employed in the classification using a Random Forest, with  $n$  assuming the values 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 140, 150.

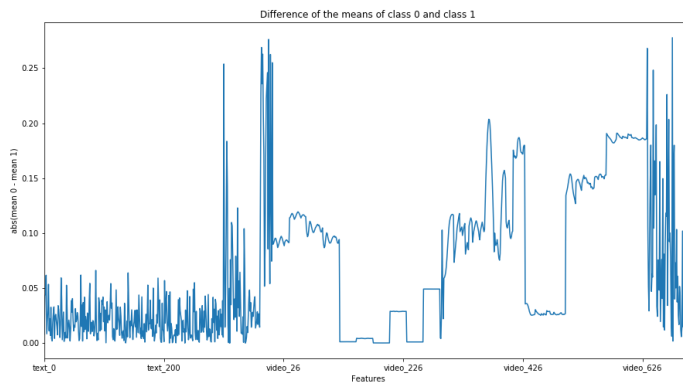


Figure 5.24: Difference of the means of the features from class 0 and class 1

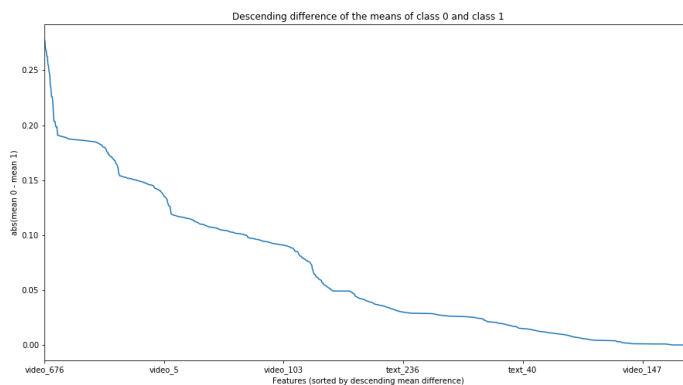


Figure 5.25: Difference of the means of the features from class 0 and class 1 sorted in descending order

A Random Forest is an ensembling method based on a set of decision tree, that operates on different batches of the data and with different features. Bootstrap aggregating with feature bagging makes the training more resistant to overfitting and improves the general accuracy.

The algorithm was iterated 20 times to get a more accurate expectation of the results and varying the number of trees in the forest (10, 50 or 100). This method resulted in a general best accuracy of 77.42%, achievable using 100 trees and 130 features (see Figure 5.27), achieved on the validation set. This is considered a good result, but it remains to be verified whether the Memory Fusion Network will outperform it. More results of the classification with the Random Forest, including precision, recall and F-score for both classes, are available in Figures 5.30a, 5.30b, 5.30c, 5.30d, 5.28e and 5.28f. Finally, table 5.1 shows the scores obtained by the best performing Random Forest algorithm on the validation and test set.

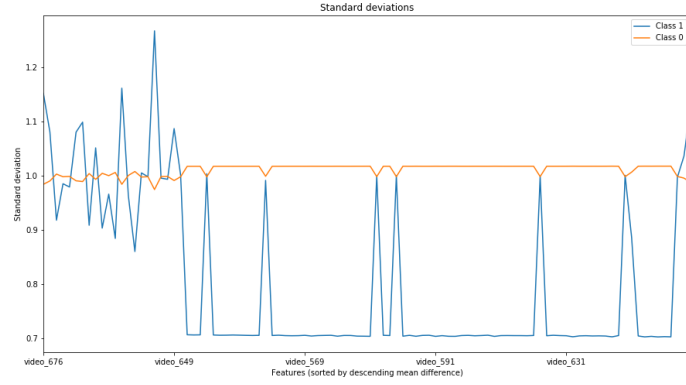


Figure 5.26: Standard deviations of the features from class 0 and class 1, sorted according the descending order of the means difference

Metric	Apparent error/accuracy		True error/accuracy	
	Class 0	Class 1	Class 0	Class 1
Accuracy		0.7672		0.6235
Precision	0.7764	0.7586	0.9312	0.0762
Recall	0.7506	0.7838	0.6419	0.3839
F-Score	0.7633	0.7710	0.7600	0.1272

Table 5.1: Accuracy obtained employing the Random Forest algorithm, trained including the 130 most discriminating features and 100 trees.

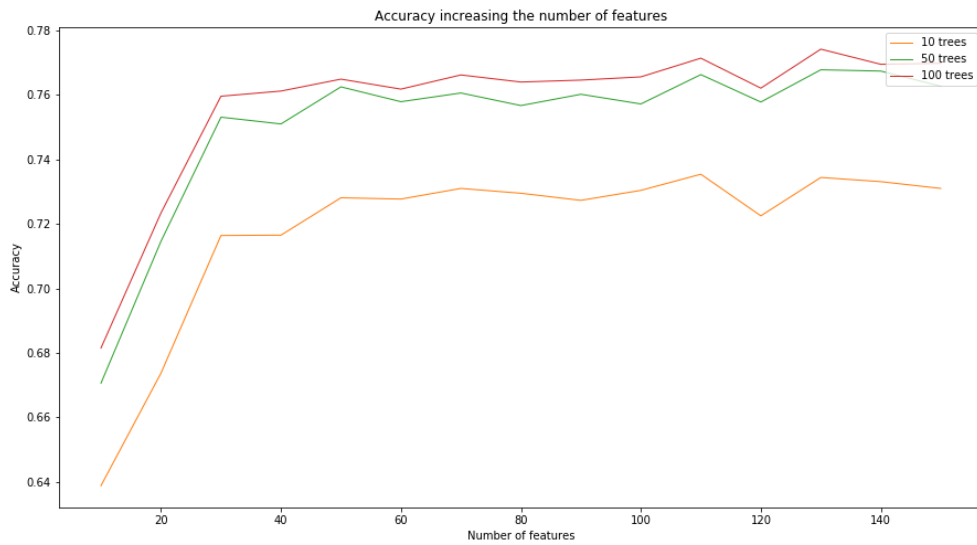
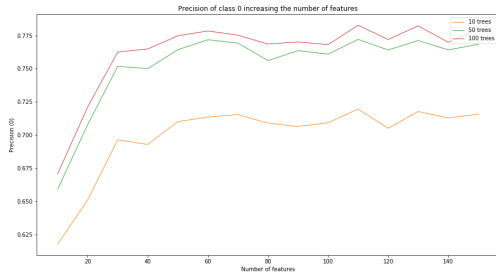


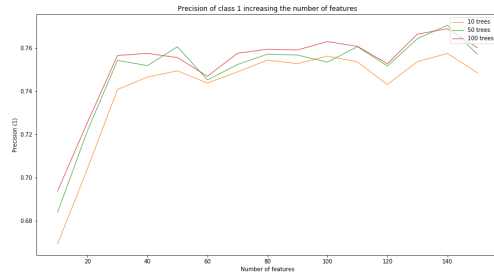
Figure 5.27: Accuracy of Random Forest with different number of trees and features

## 5.5. MACHINE LEARNING MODELS WITH MULTIMODAL FEATURES INTEGRATION

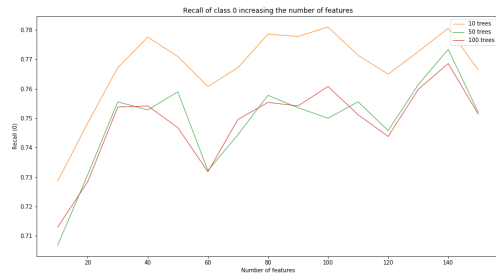
This section described the second approach for multimodal Machine Learning, based on feature integration through the Memory Fusion Network architecture from [12]. The implementation of the Memory Fusion Net-



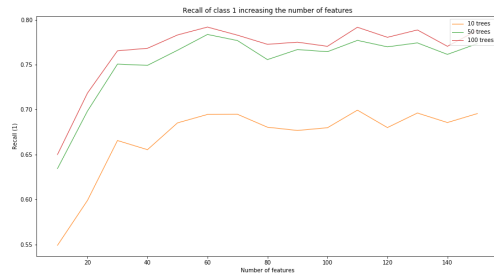
(a) Random forest: precision for class 0



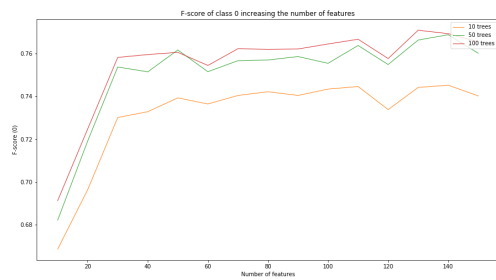
(b) Random forest: precision for class 1



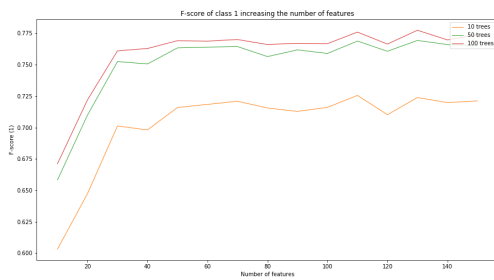
(c) Random forest: recall for class 0



(d) Random forest: recall for class 1



(e) Random forest: F-score for class 0



(f) Random forest: F-score for class 1

work, shared by the authors, is publicly available on GitHub<sup>5</sup>. The code, written in Python 2 using the library PyTorch<sup>6</sup>, was made compatible for Python 3 and adapted to the problem of video segment classification based on the segments saliency, which is tackled in this thesis. By definition, a segment of a political speech is considered relevant if it is likely to be included in a highlight clip video from a news channel.

### 5.5.1. INPUT DATA

The input data of the MFN is a three dimensional matrix, visible in Figure 5.29. One dimension corresponds to the video segments: for each segment, the two other dimensions belong to the concatenated multimodal features and the time steps, where each time step corresponds to one pronounced word. The number features correspond to:

- 300-dimensional text features,
- 72-dimensional audio features (2 columns were removed because they had zero variance),
- 709-dimensional video features.

Since the segments have different lengths, in order to collect all of them in one matrix, the sequence length was fixed to a maximum of 100 time steps, namely 100 words. The sequences that are shorter are padded

<sup>5</sup><https://github.com/pliang279/MFN>

<sup>6</sup><https://pytorch.org>

with zeros, starting from the first row, so to obtain right alignment, while longer sequences are truncated. The value 100 for the maximum length was chosen in order not to cut too many segments, since only 5% are longer than 100, but at the same time not to slow down the computation by adding a too many dimensions to the 3D matrix.

The dataset is divided in a way to maintain 50 speeches in the training set, 20 in the validation set and 29 in the test set. The test set is fixed, and contains a total of 5564 video segments, while speeches for the validation set are randomly sampled every time the code is executed. Training set and validation set contain 10060 video segments in total.

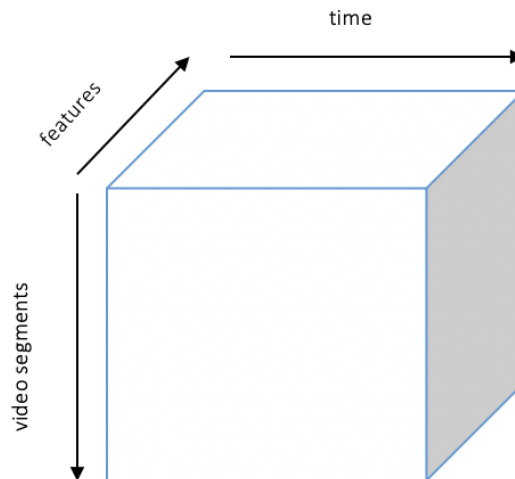


Figure 5.29: Structure of the input of the MFN

### 5.5.2. MODIFICATIONS IN THE CODE

Following the poor results of some preliminary attempts, the implementation of the MFN was modified in order to take into account the strong imbalance in the classes of the dataset (93% of segments belong to class 0 and only 7% to class 1).

The original implementation simply uses L1 loss to measure the performance of the network and train it accordingly. The L1 loss measures the Mean Absolute Error (MAE), defined as

$$l(x, y) = L = \{l_1, \dots, l_N\}^T, \quad l_n = |x_n - y_n|.$$

When the neural network is trained with the L1 loss, it learns to assign all the data points to class 0, this way always obtaining high accuracy on the mini batches (above 90%). In order to balance the predictions, the L1 loss was substituted with the Cross Entropy Loss, which allows to specify a weight for each class:

$$\text{loss}(x, \text{class}) = \text{weight}[\text{class}] \left( -x[\text{class}] + \log \left( \sum_j \exp(x[j]) \right) \right).$$

The weights were chosen to be the inverse of the number of elements per class, namely  $\frac{1}{93}$  for class 0 and  $\frac{1}{7}$  for class 1. The weights are automatically normalised by the loss function to sum to 1.

The Cross Entropy Loss receives as input the ground-truth labels and a matrix whose dimensions corresponds to the number of data points (the batch size during training) and the number of classes, two in this case. Each cell is supposed to correspond to the predicted probability of a data to be assigned to class 0 or class 1. Since the original MFN was implemented in a way to produce just one scalar number, the last layer was modified in order to output two values. However, it was not necessary to add a softmax activation to the output layer, to constrain the numbers to range from 0 to 1 and sum to 1, because the PyTorch implementation of the Cross Entropy Loss already includes LogSoftmax in combination with the negative log likelihood loss<sup>7</sup>.

<sup>7</sup><https://pytorch.org/docs/stable/nn.html#crossentropyloss>

The effect of the Cross Entropy Loss is to penalise wrong predictions, with more attention to class 1, but also to penalise correct predictions made with low confidence, thus making the predictions accurate with more probability. This results in higher precision/recall values [138].

While training, the model that improves the current best loss on the validation set is saved.

### 5.5.3. HYPERPARAMETERS

The configurations of the original implementation were preserved. Here is a list of the hyperparameters used during training.

- Output dimensions of the systems of LSTMs
  - $h_{text} = \text{random\_choice}(32,64,88,128,156,256)$
  - $h_{audio} = \text{random\_choice}(8,16,32,48,64,80)$
  - $h_{video} = \text{random\_choice}(8,16,32,48,64,80)$
- Multi-view Gated Memory dimension
  - $memsize = \text{random\_choice}(64,128,256,300,400)$
- Window size for the DMAN
  - $windowsize = 2$
- Training options
  - $batchsize = \text{random\_choice}(32,64,128,256)$
  - $num\_epochs = 50$
- Optimizer parameters
  - $learning\_rate = \text{random\_choice}(0.001,0.002,0.005,0.008,0.01)$
  - $momentum = \text{random\_choice}(0.1,0.3,0.5,0.6,0.8,0.9)$
- Network specifics for the DMAN
  - $dimensions = \text{random\_choice}(32,64,128,256)$
  - $dropout = \text{random\_choice}(0.0,0.2,0.5,0.7)$

### 5.5.4. FIRST RESULTS

The MFN was trained several times tweaking the hyperparameters described in the previous section. Table 5.2 reports some results obtained by five different models, trained simply using the L1 loss function. The table shows the classification results on the validation set. Clearly, the results are poor. A possible explanation for these results is that the MFN is not able to capture the common characteristics of relevant segments. In fact, the low level feature used for the classification might be insufficient for such a high-level task. Another possibility is that using the L1 loss for the evaluation on the validation set is not a suitable metric: even though, while training, the Cross Entropy Loss takes into account the classes imbalance, the L1 loss does not. In order to take into account the strong data imbalance, the network is trained again on an augmented version of the dataset and using the Cross Entropy Loss.

### 5.5.5. DATA AUGMENTATION

The technique of data augmentation consists in generating multiple new samples from the same distribution of the target data, so that it becomes easier to analyse it [139]. Considering the Political Speeches Dataset, data augmentation can be used to create new instances of data points labelled 1, in order to compensate the dataset unbalance and simplify the difficulty of the MFN in modelling the data distribution.

The data augmentation process was performed by creating new data points, starting from the samples of the training set labelled 1 and adding to the corresponding feature vectors Gaussian noise, with  $\mu = 0$  and  $\sigma = 0.1$ . The number of new samples was selected such that the augmented datasets contained approximately

Model	Metric	Apparent error/accuracy	
		Class 0	Class 1
<i>mfn_67744</i>	Mean Absolute Error		0.0971
	Precision	0.9297	0.0747
	Recall	0.9689	0.0332
	F-score	0.9489	0.0459
<i>mfn_57501</i>	Mean Absolute Error		0.1010
	Precision	0.9307	0.0991
	Recall	0.9631	0.0536
	F-score	0.9466	0.0695
<i>mfn_43357</i>	Mean Absolute Error		0.1051
	Precision	0.9293	0.0633
	Recall	0.9600	0.0357
	F-score	0.9444	0.0457
<i>mfn_52167</i>	Mean Absolute Error		0.1229
	Precision	0.9305	0.0852
	Recall	0.9377	0.0765
	F-score	0.9341	0.0806
<i>mfn_5858</i>	Mean Absolute Error		0.0902
	Precision	0.9297	0.0769
	Recall	0.9768	0.0255
	F-score	0.9527	0.0383

Table 5.2: Results of the different MFN models

50% of data points from both classes. Several models were trained with different hyperparameters and using the Cross Entropy Loss, with class weights  $(\frac{1}{2}, \frac{1}{2})$ . Table 7.2 shows the results obtained by four different models, trained on the augmented dataset, after 30 epochs.

Comparing Table 5.2 and 5.3, it can be noticed that the use of Cross Entropy loss and data augmentation results in a clear improvement over simple training with the L1 loss and without managing the data unbalance. However, even if the accuracy achieved on the validation set during training significantly improves, the accuracy on the training set is very low. In particular, the chosen best performing model (*mfn\_aug\_bs64*, trained with batch size 64) only scores approximately 8% or precision and recall for class 1, corresponding to the salient segments. The disparity between the apparent accuracy (measured on the validation set) and the true accuracy (measured on the test set), means that the models are overfitting the training set. Overfitting is confirmed by the plot of the Cross Entropy Loss during the training phase, visible in Figure 5.30. As can be seen in the plots, the models loss on the training set rapidly approaches zero during the first five epochs. On the other hand the validation losses have an oscillating trend, but overall tend to increase over time. In order to prevent the Cross Entropy Loss from growing too much, early stopping is applied by limiting the epochs to 30. Besides data augmentation and early stopping, also dropout is used to prevent overfitting. However, these expedients are still not sufficient to obtain better accuracy. The incapacity of the models to generalise over new data samples might be due to the fact that the speeches content are too different from one another to be able to construct a single supervised model that extracts highlights from all the possible speeches. In addition, the presence of different speakers in the videos to classify increases the difficulty of the problem.

In order to reduce the complexity of the classification, additional MFN models were trained and tested only on speeches from the same politician, either Barack Obama or Donald Trump. However, this attempt did not result in any improvements.

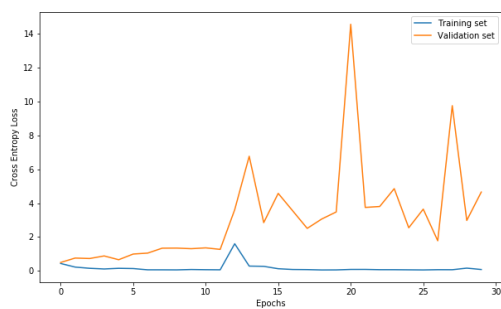
In spite of the low precision and recall scores obtained on the test set for class 1, the chosen best performing MFN model, *mfn\_aug\_bs64*, was used to extract the video segments candidate for the highlight clip. The model was applied to the 29 speeches of the test set. For each of them, it classified as salient at least 1 segments. Out of 29 speeches, for 20 speeches the amount of segments classified as salient ranges from 1 to 10. The video segments classified as salient are concatenated together in their chronological order to form the "MFN highlight clips".

Even though the MFN highlight clips do not share many video segments with the ground truth, hopefully the MFN architecture is able to capture some properties of the highlight video segments that can be used as a criteria to detect highlights. These characteristics can correspond to the presence of a particular sentence

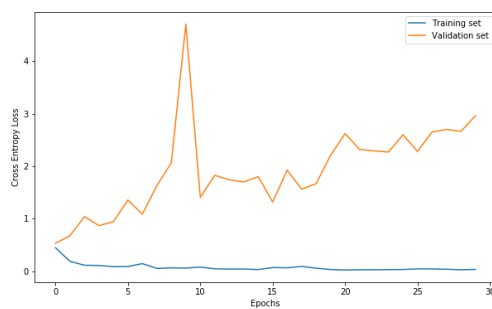
Model	Metric	Apparent error/accuracy		True error/accuracy	
		Class 0	Class 1	Class 0	Class 1
<i>mfn_aug_bs32</i>	Mean Absolute Error		0.4987		0.0626
	Precision	0.4843	0.9954	0.9377	0.0000
	Recall	0.9997	0.0584	0.9996	0.0000
	F-score	0.6525	0.1104	0.9677	0.0000
<i>mfn_aug_bs64</i>	Mean Absolute Error		0.0023		0.1159
	Precision	0.9992	0.9973	0.9386	0.0753
	Recall	0.9969	0.9993	0.9386	0.0753
	F-score	0.9981	0.9983	0.9386	0.0753
<i>mfn_aug_bs128</i>	Mean Absolute Error		0.0000		0.1204
	Precision	1.0000	1.0000	0.9368	0.0481
	Recall	1.0000	1.0000	0.9365	0.0484
	F-score	1.0000	1.0000	0.9366	0.0483
<i>mfn_aug_bs256</i>	Mean Absolute Error		0.0000		0.0796
	Precision	1.0000	1.0000	0.9384	0.0923
	Recall	1.0000	1.0000	0.9789	0.0323
	F-score	1.0000	1.0000	0.9582	0.0478

Table 5.3: Results of the best performing MFN model trained on the augmented dataset

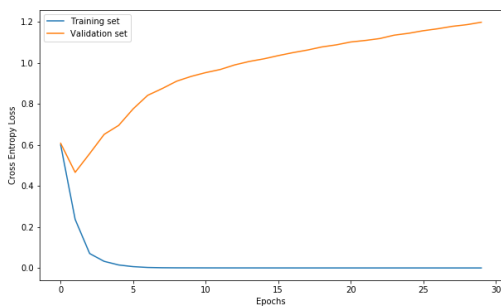
in the speech, a certain facial expression or tone of the voice. Finally, the quality of the MFN highlight clips is assessed through crowdsourcing. Additional information will follow in Chapter 6 and Chapter 7.



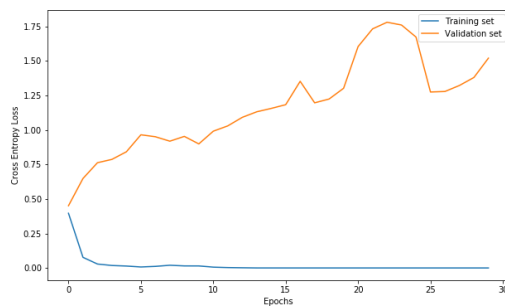
(a) Batch size = 32



(b) Batch size = 64



(c) Batch size = 128



(d) Batch size = 264

Figure 5.30: Cross Entropy Loss during training





# 6

## CROWDSOURCING METHODOLOGY

### 6.1. INTRODUCTION

A high level evaluation of the thesis results is conducted through a crowdsourcing user study, using the platform Figure Eight<sup>1</sup>. This chapter describes into detail the design and implementation of the crowdsourcing process.

The crowdsourcing tasks are designed in order to be able to assess the performance of the different methods for the creation of automatic highlight clips explored in the thesis, and to compare them against the ground truth highlights. The highlight clips types assessed through crowdsourcing include:

1. ground truth highlights;
2. highlights created with MFN;
3. highlights created with the Random Forest algorithm;
4. baseline highlights, extracted from the transcripts summarisation.

All the highlight clips presented in the crowdsourcing tasks are extracted from the 29 speeches that belong to the test set, thus obtaining a fair evaluation of the Machine Learning methods involved.

Before publishing the final crowdsourcing jobs, a pilot evaluation round was performed in order to estimate the clearness and effectiveness of the tasks, the number of workers required and the budget. The preliminary results from the pilot evaluations are included in the following sections.

### 6.2. EVALUATION WITH CROWDSOURCING

The design of the crowdsourcing tasks employed for the evaluation of the thesis results begins with the definition of the basic crowdsourcing elements identified in [110], namely:

1. Crowd
2. Initiator
3. Process

The following subsections are dedicated to the deepening of these three, in relation to the objective of the thesis. Since the crowdsourcing process is conducted in two rounds, first a pilot batch and then the final evaluation, the elements concerning these two steps will be referred to as *A* and *B*.

---

<sup>1</sup><https://www.figure-eight.com>

### 6.2.1. CROWD

- Who forms the crowd?

A Group of 30 people who completed at least one task from the pilot evaluation round. They include: interns and supervisors at IBM CAS, fellow students from TU Delft, relatives and friends, additional friends recruited through Facebook, all with full proficiency in English. This group was distributed in such a way that each task was completed by 10 different people. The results are used to estimate the effectiveness of the questions, the required number of people per task and the hourly pay.

B Professional crowdworkers from Figure Eight [134]. Since the videos from the datasets are in English, the crowd must be formed of English speakers, therefore people from Australia, Canada, Ireland, New Zealand, United Kingdom and United States are chosen.

- What does the crowd have to do?

Two surveys are available on Figure Eight [134]: an *individual assessment* of one single highlight clip at a time and a *pairwise comparison*, where two different highlight clips from the same speech are compared. The first survey is used to obtain the evaluation of each highlight clip, independently on the others, while in the second survey the workers are forced to choose the best out of two highlight clips, according to different criteria. A ranking of the different methodologies for the highlights creation will be derived from the results of the second survey. Each worker can choose to fill one of the two, or both. In the first survey, which will be referred to as *individual assessment*, the workers are required to watch one of the highlight clips generated with the methods mentioned in the introduction §6.1, answer two Likert questions and two *Yes/No* questions. The workers must also motivate their choice, which forces them to reflect on the questions and to provide sensible answers. One highlight clip, plus the respective questions, form one task. The survey contains the following questions and possible answers:

**Q1** *This highlight clip gives me information about what the full speech was about.* How much do you agree with this statement?

- a Strongly disagree
- b Disagree
- c Neither agree nor disagree
- d Agree
- e Strongly agree

**Q2** *This highlight clip makes me want to watch the full speech.* How much do you agree with this statement?

- a Strongly disagree
- b Disagree
- c Neither agree nor disagree
- d Agree
- e Strongly agree

**Q3** I like the politician in the video.

- a Yes
- b No

**Q4** I have already heard this speech.

- a Yes
- b No

**Q5** Did you find it difficult to answer these questions? Why or why not?

Question Q1 and Q2 are used to measure the ability of the highlights to convey the message of the original speech and generate interest in the viewers, while Q3 and Q4 serve to identify correlations between eventual political biases or prior knowledge about the speech and the answers to Q1 and Q2. Question Q5 provides the crowdworkers feedback about the surveys understandability and feasibility, especially useful for the crowdsourcing tasks of type A.

The second survey, the *pairwise comparison*, contains five main double-choice questions. The workers are required to watch two highlight clips and then answer the questions, choosing either the first highlight clip or the second highlight clip. In addition, they must provide the motivation to their answers and provide their feedback about the difficulty of the survey questions. One task from this survey involves exactly one highlight clips pair and the respective questions. The questions are:

- Q1 Which video gives more information about the topic of the full speech?
- Q2 Which video contains more repetitions of the same sentences?
- Q3 Which video is more engaging (independently on its content or the political tendency)?
- Q4 Which video is more cinematic?
- Q5 In which video is the speaker more expressive?
- Q6 Did you find it difficult to answer these questions? Why or why not?

The questions from Q1 to Q5 are used to extract information about the degree of informativeness, the level of entertainment, the presence of repetitions, the expressiveness of the speaker, which together contribute to defining the relative quality of the different highlight clips.

An example of what the surveys on Figure Eight look like is given by Figure B.1 for the individual assessment and Figure B.2 for the pairwise comparison.

- What does the crowd get in return?
  - A No monetary recompense is contemplated for the first group of workers, who contributed in the pilot evaluation round.
  - B The professional crowdworkers receive monetary compensation for every survey completed. The precise amount is calculated based on the effort and the time needed to complete the tasks thoroughly. These are estimated from group A. The price per task is to be calculated with respect to the Frankfurt Declaration on Platform-Based Work<sup>2</sup>, for an equivalent of \$7.25 per hour, the minimum wage in the United States [140].

### 6.2.2. INITIATOR

- Who is the initiator?
 

Ombretta Strafforello, graduating MSc student at TU Delft.
- What does the initiator get in return?
 

The initiator obtains a complete assessment of the highlight clips produced for the present thesis. From these assessment it is possible to evaluate the extent of effectiveness of the methods for automatic highlight clips generation introduced in the thesis.

### 6.2.3. PROCESS

- What type of process is it?
 

It is a distributed outsourcing process on the internet, exploiting the Figure Eight crowdsourcing platform.
- What type of call to use?
  - A The first group of workers are directly asked to collaborate through online messages or through a Facebook post, through which the surveys links are shared.
  - B The surveys are uploaded on Figure Eight [134], where are visible by the professional crowdworkers, who can freely choose to collaborate.
- Which medium is used?
  - A Direct connection with colleagues, relatives and friends via internet, using online messages services and social networks, like Facebook.
  - B Connection with the professional crowdworkers via internet, through the Figure Eight [134] crowdsourcing platform.

<sup>2</sup><http://faircrowd.work/unions-for-crowdworkers/frankfurt-declaration/>

### 6.3. PILOT SURVEYS

The pilot surveys are designed according to the approach described in the previous section, §6.2, and they consist of an individual assessment and a pairwise comparison of different highlight clips. They are used to understand whether the answers to the composed questions can be effectively used to obtain an objective evaluation of the methods for automatic highlights generation, to estimate the required number of crowdworkers per task and their hourly pay.

The pilot surveys contain a small batch of tasks, namely 10 highlight clips in the *individual assessment* survey and 5 highlight clip pairs from the *pairwise assessment* survey, therefore 10 tasks from the first survey and 5 from the second. Only ground truth highlights and highlights generated with MFN are present in the pilot, all corresponding to speeches from the test set. The highlight clips were presented in random order, to avoid a possible bias in the answers. Thirty workers were recruited out of colleagues, family and friends, and each task was completed by 10 different workers. The surveys were active from July 31st, 2019, to August 5th, 2019. The results and considerations from the pilot surveys are reported in the following sections.

### 6.4. ESTIMATION OF THE NUMBER OF CROWDWORKERS REQUIRED

The estimation of the number of crowdworkers required for the jobs is performed following the same methodology as the paper titled "How many crowdsourced workers should a requester hire?" [141]. The authors of the paper present a precise analysis of the minimum required number of crowdworkers per task, in order to obtain consistent answers when aggregating the reported outputs.

In order to do so, the authors designed three evaluation tasks and assigned 50 recruited crowdworkers to complete each task. The accuracy of the aggregate outputs in terms of the number of workers was estimated by iterating a bootstrap resamples procedure 100,000 times for each number of workers  $n \in \{1, \dots, 50\}$ . At each iteration, the aggregate output is calculated by averaging the answers of the  $n$  sampled workers. The aggregated output is then compared to the "gold-standard output", defined as the mode of the outputs by all the 50 workers. The difference between the gold-standard output and the aggregated output is measured in terms of mean squared error (MSE). The mean MSE and the respective standard deviation are calculated out of the 100,000 iterations and are used to express the expected error. Naturally, the larger the number of workers  $n$ , the smaller the resulting expected error. If  $n$  is large enough, the MSE becomes approximately constant. The objective is to identify where the constant region starts, so that it can be concluded that only  $n \leq 50$  workers are needed to obtain an accurate aggregated output. This is achieved by defining the threshold under which the MSE is considered stable.

The threshold is derived by means of a piecewise linear regression analysis. Namely, the MSE curve is approximated using a series of line segments. The last line segment corresponds to the constant region, therefore the breakpoint that separates the last two line segments is the starting point where the average error can be considered stable. The authors of [141] rely on the dynamic programming algorithm proposed in [142] for the calculation of the optimal number of line segments to use in the approximation, which results to be 3. The same number is adopted for the analysis of the present crowdsourcing process. The results obtained for the two surveys and additional details about the calculations are reported as follows.

#### 6.4.1. INDIVIDUAL ASSESSMENT

The accuracy of the aggregate output was estimated for the first survey, the *individual assessment*, for the two rating scale questions:

- Q1: "This highlight clip gives me information about what the full speech was about. How much do you agree with this statement?"
- Q2: "This highlight clip makes me want to watch the full speech. How much do you agree with this statement?"

The five possible answers to the questions, namely:

- a Strongly disagree
- b Disagree
- c Neither agree nor disagree
- d Agree

e Strongly agree

are encoded as, respectively, -2,-1,0,1,2, to allow the calculations of the MSE. In such a way, the scale is symmetrical and a negative score is associated to a negative worker response.

The gold-standard output for the survey, that is the mode of all the answers for each of the 10 tasks, is:

	Task number									
	1	2	3	4	5	6	7	8	9	10
<b>Q1</b>	-2	1	1	0	0	1	2	2	0	1
<b>Q2</b>	-2	-1	0	-1	2	-1	0	0	1	0

Table 6.1: Gold-standard output for the individual assessment

Figure 6.1 and 6.2 report the MSE between the gold-standard output and the aggregated answer in terms of the number of workers considered. The plot on the left shows the mean and standard deviation, estimated after iterating 1000 bootstrap resamples, while the plot on the right shows the piecewise linear approximation of MSE using three segments.

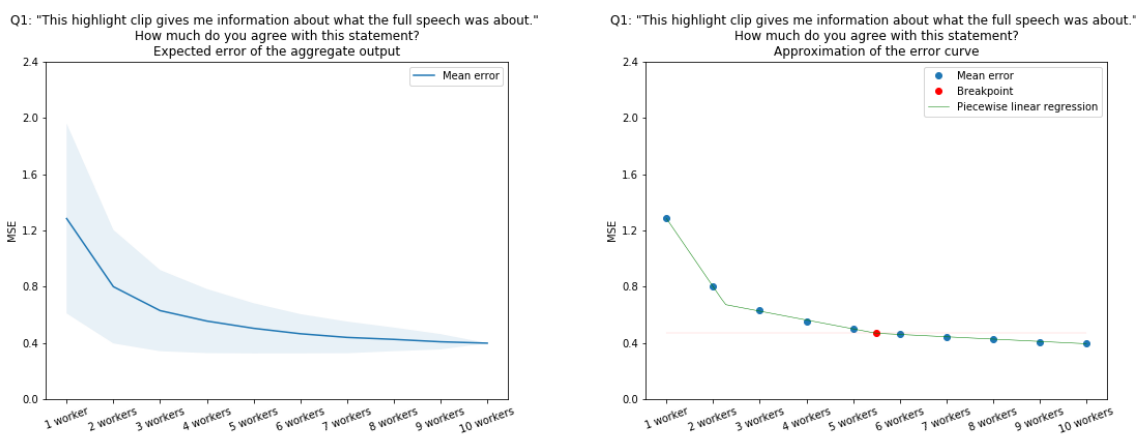


Figure 6.1: Q1: "This highlight clip gives me information about what the full speech was about. How much do you agree with this statement?"

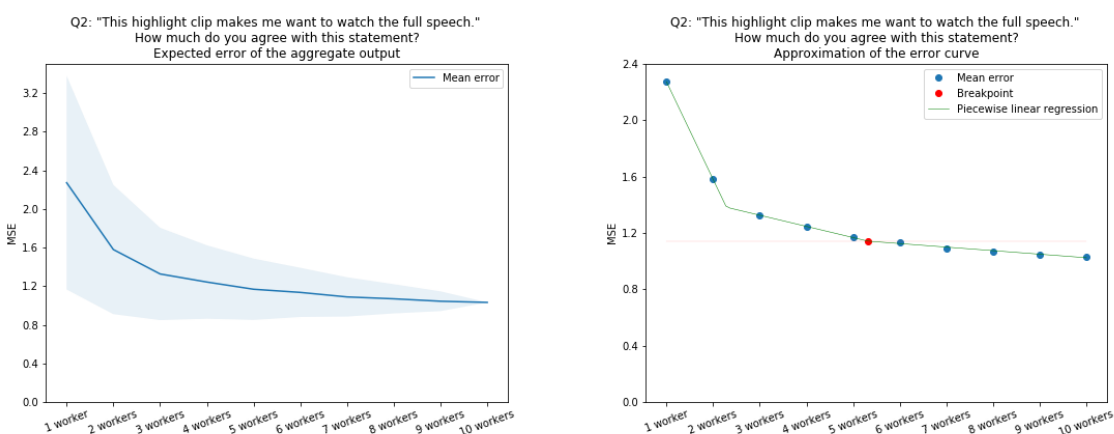


Figure 6.2: Q2: "This highlight clip makes me want to watch the full speech. How much do you agree with this statement?"

According to the position of the breakpoint (red dot in the plots), which is collocated at  $n = 5.50$  for Q1

and  $n = 5.32$  for Q2, it can be concluded that at least 6 workers are necessary for each task of the *individual assessment*.

#### 6.4.2. PAIRWISE COMPARISON

A similar analysis was conducted for the second survey, in particular for the five questions:

- Q1: "Which video gives more information about the topic of the full speech?"
- Q2: "Which video contains more repetitions of the same sentences?"
- Q3: "Which video is more engaging (independently on its content or the political tendency)?"
- Q4: "Which video is more cinematic?"
- Q5: "In which video is the speaker more expressive?"

The possible answers to the questions, that is

- First highlight clip
- Second highlight clip

were encoded as 0 and 1, respectively. The gold-standard output for the *pairwise comparison* was derived from the mode of the answers by all 10 workers and resulted to be:

Task number	Task number				
	1	2	3	4	5
<b>Q1</b>	0	0	0	1	0
<b>Q2</b>	0	0	1	1	1
<b>Q3</b>	0	0	1	0	0
<b>Q4</b>	1	0	1	0	0
<b>Q5</b>	0	0	1	0	0

Table 6.2: Gold-standard output for the pairwise comparison

Figure 6.1 and 6.2 report the MSE between the gold-standard output and the aggregated answer in terms of the number of workers considered. The plot on the left shows the mean and standard deviation, estimated after iterating 1000 bootstrap resamples, while the plot on the right shows the piecewise linear approximation of MSE using three segments.

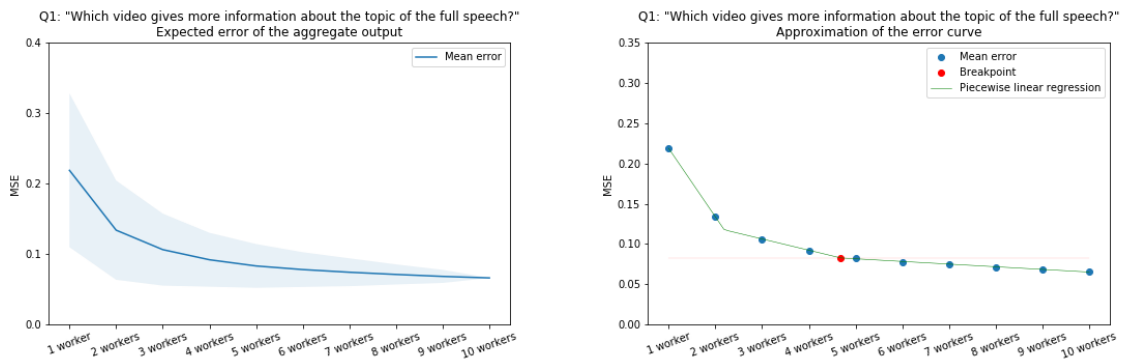


Figure 6.3: Q1: "Which video gives more information about the topic of the full speech?"

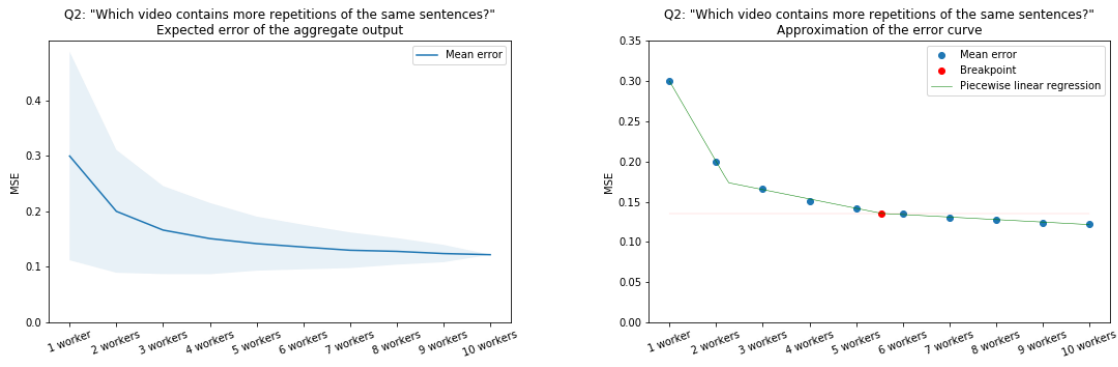


Figure 6.4: Q2: "Which video contains more repetitions of the same sentences?"

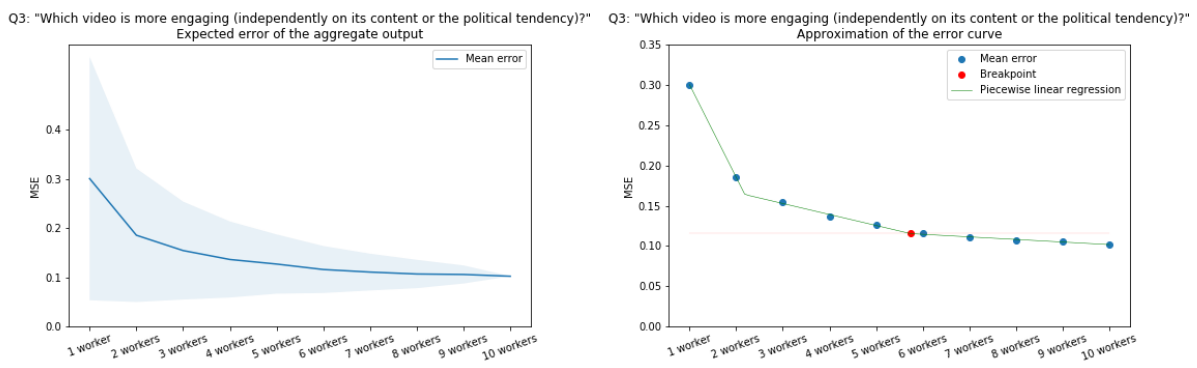


Figure 6.5: Q3: "Which video is more engaging (independently on its content or the political tendency)?"

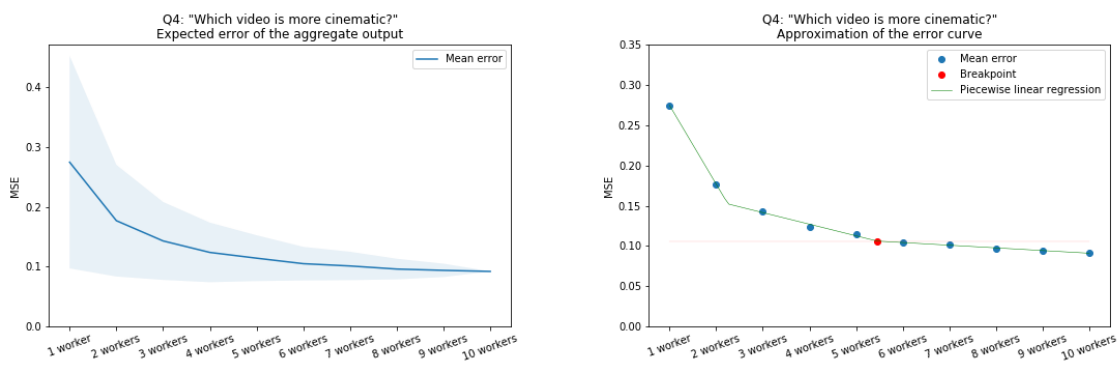


Figure 6.6: Q4: "Which video is more cinematic?"

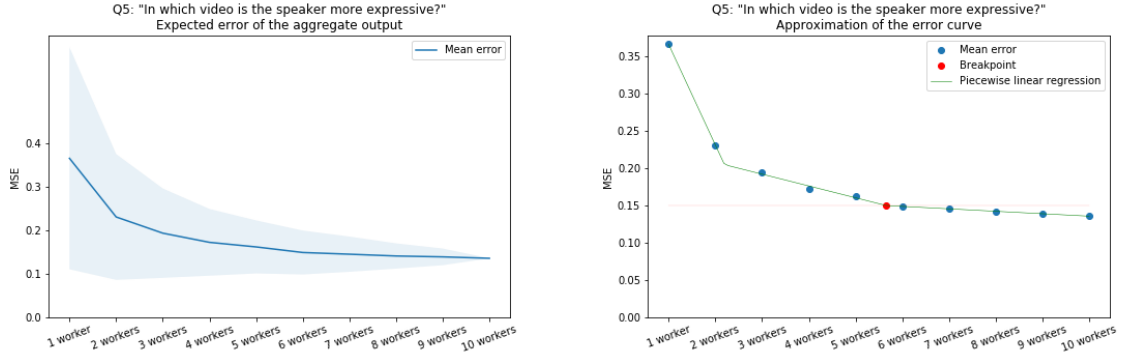


Figure 6.7: Q5: "In which video is the speaker more expressive?"

The position of the breakpoints for the five questions are: 4.68, 5.60, 5.74, 5.45 and 5.66, therefore it can be concluded that at least 6 workers are required for each task of the *pairwise comparison*.

## 6.5. ESTIMATION OF THE BUDGET

The calculation of the budget is based on a hourly pay of \$7.25 for the crowdworkers, which is the minimum wage in the United States [140]. The respective pay per task is derived taking into consideration the average time the crowdworkers spent on the tasks from the two surveys.

If the four types of highlight clips generated for all the 29 speeches from the test set were included in the two surveys, the total estimated budget would be \$1774.80. Due to the prohibitively expensive cost, the clips for only 5 speeches are considered. This way, the required budget is reduced to \$306.

All the details regarding the calculations are reported as follows.

### 6.5.1. INDIVIDUAL ASSESSMENT

According to the Figure Eight report generated for the results and statistics of the pilot surveys, a crowdworker spent 5m 1s for each task from the *individual assessment* survey, on average. Therefore, a worker can complete approximately 12 tasks per hour. In order to guarantee a hourly pay of at least \$7.25, the price per task should be  $\$7.25/12 \approx \$0.60$ .

The *individual assessment* survey contains four tasks for each of the 29 speeches from the test set, namely one task for every method for highlight generation. This corresponds to a total of 116 tasks. Allowing 6 crowdworkers per task, the total budget required for the *individual assessment* survey is:

$$116 \times 6 \times \$0.60 = \$417.60.$$

If only 5 speeches are considered, the total tasks become 20, which corresponds to a cost of:

$$20 \times 6 \times \$0.60 = \$72.00.$$

### 6.5.2. PAIRWISE COMPARISON

The budget for the *pairwise comparison* survey is calculated similarly as for the *individual assessment* survey. According to what reported by Figure Eight, a crowdworker spent on average 10m 27s on each task from the survey. This corresponds to an estimated hourly pay of approximately  $\$7.25/(60/10.5) = \$1.30$ .

Each survey contains one task for each highlight clip pair, in such a way that all the four highlight clips generated for one speech are compared against each other. This adds up to  $\binom{4}{2} = 6$  highlight clips pair per speech. Since there are 29 speeches in the test set, a total of 174 tasks are contained in the survey. Considering the previously calculated hourly pay and requiring exactly 6 workers per task, the total budget is:

$$174 \times 6 \times \$1.30 = \$1357.20.$$

If only 5 speeches are considered, the total tasks become 30, which corresponds to a cost of:

$$30 \times 6 \times \$1.30 = \$234.00.$$



## 6.6. PRELIMINARY RESULTS

The two plots in Figure 6.8 show the mean output, obtained for the *individual assessment* pilot survey. In the plots, the tasks corresponding to the ground truth highlights are colored in red, while the ones corresponding to the highlights generated with MFN are in green. At a first glance, it can be noticed that, overall, the ground truth highlights are marked by a greater ability in conveying the message of the original speech and spark interest in the viewers. However, this trends needs to be confirmed by the final crowdsourcing round, which will include also the highlights created with the Random Forest algorithm and the highlights extracted from the transcripts summarisation.

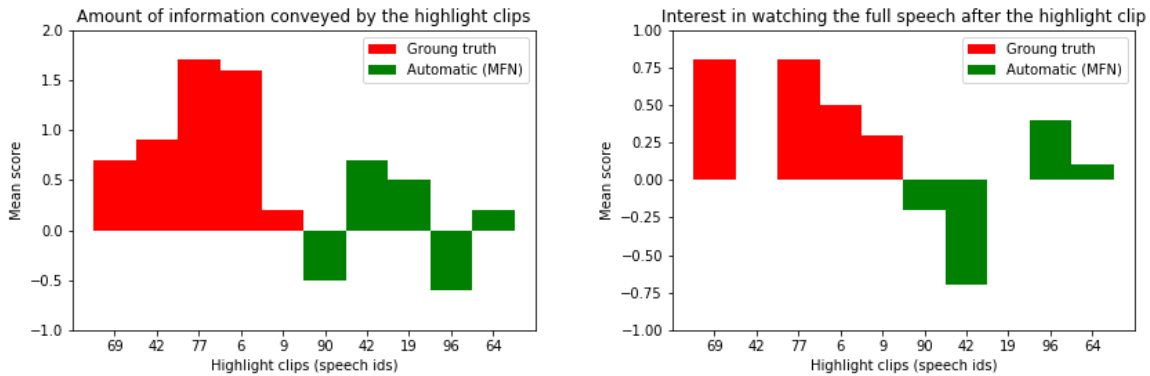


Figure 6.8: Mean output for the *individual assessment* survey

Also for what concerns the *pairwise comparison* survey, the ground truth highlights are viewed more favourably than the highlights from MFN. The votes for each video pair are visible in 6.9. In particular, the total aggregated results for the five main questions are as follows:

- Q1: "Which video gives more information about the topic of the full speech?"
  - Ground truth: 66%
  - MFN: 34%
- Q2: "Which video contains more repetitions of the same sentences?"
  - Ground truth: 54%
  - MFN: 46%
- Q3: "Which video is more engaging (independently on its content or the political tendency)?"
  - Ground truth: 62%
  - MFN: 38%
- Q4: "Which video is more cinematic?"
  - Ground truth: 64%
  - MFN: 36%
- Q5: "In which video is the speaker more expressive?"
  - Ground truth: 64%
  - MFN: 36%

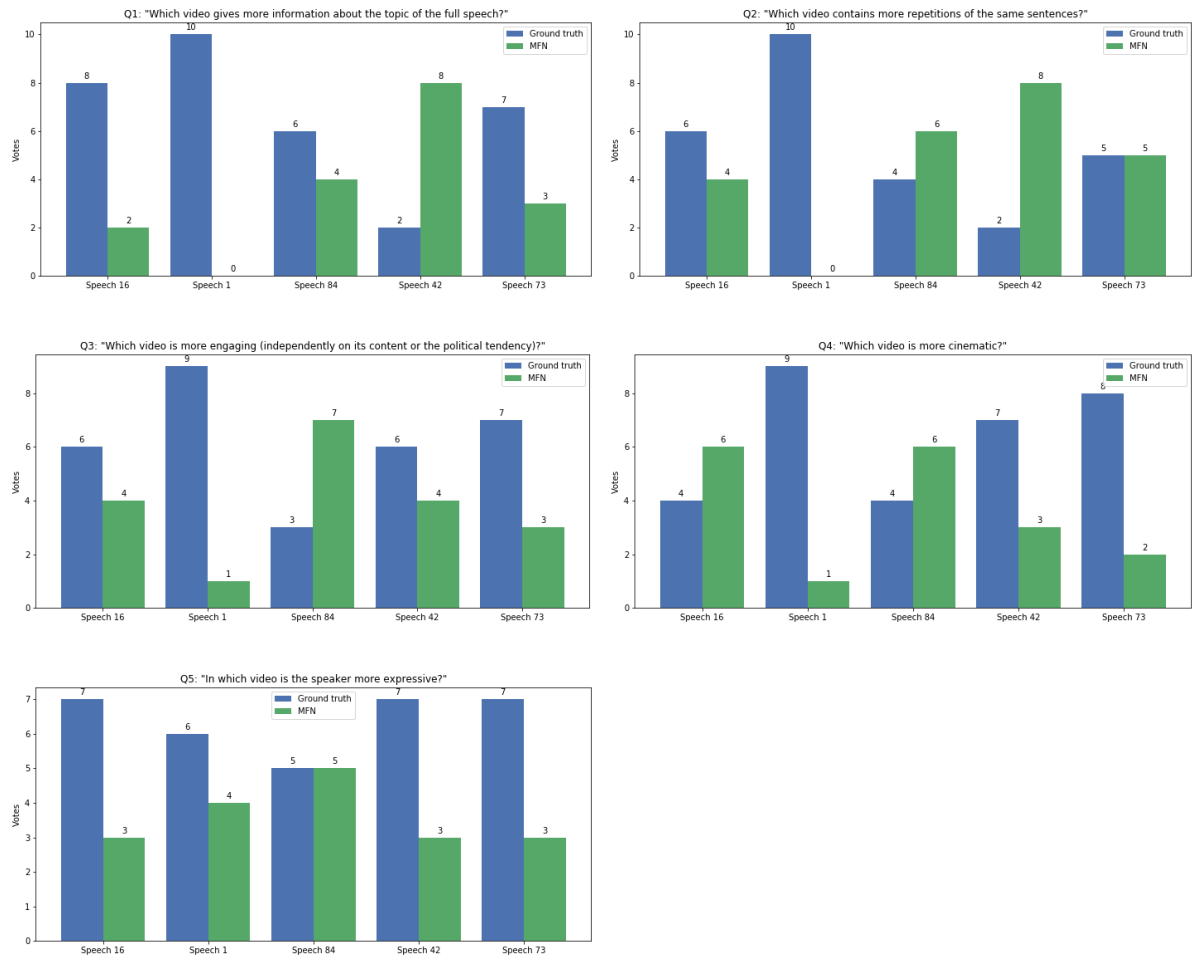


Figure 6.9: Workers votes in the *pairwise comparison* survey

# 7

## RESULTS AND DISCUSSION

### 7.1. INTRODUCTION

For the research purpose subject of this thesis, namely the exploration of multimodal Machine Learning techniques for highlights extraction from videos, several approaches and algorithms have been attempted. These were applied to the 99 videos collected in the Political Speeches Dataset and used to create highlights clips automatically. The experimentation with unimodal and multimodal Machine Learning techniques resulted in the creation of three different types of highlight clips for every political speech from the dataset, in addition to the ground truth highlight clips, replicated from the already existing clips made by professional video makers and broadcast by news channels.

The quality of the automatically generated highlight clips is compared to the quality of the ground truth highlights through two approaches. With respect to a highlight clip, the term quality is used to measure its resemblance to the highlight clips produced by real news channels, degree of informativeness, ability to spark interest in the viewers, level of entertainment produced and the expressiveness of the speaker.

If the quality of ground truth is similar to the quality of the automatically generated highlight clips, it means that the thesis successfully resulted in the creation of an effective automated approach based on Machine Learning. This imply a positive answer to **RQ1** and **RQ2** and the verification of the research hypothesis (**H<sub>p</sub>**: *Multimodal features can be used to train Machine Learning models that give better results in information extraction from videos, compared against unimodal methods.*).

Firstly, a low level evaluation can be directly performed, based on the calculation of precision/recall scores, that measure the amount of scenes from the original videos that are maintained in the automatic highlights and the amount of spurious scenes present in the automatic highlights. This basic method gives a measure to the ability of the algorithms to identify the scenes that also appear in the ground truth highlights, but the extent to which the method is capable of overcome the semantic gap by really understanding the linking properties and the key content of the salient scenes is something that is not possible to measure through an automatic evaluation. Therefore, the second evaluation is performed involving humans, in particular crowdworkers from the platform Figure Eight [134].

This chapter presents the results obtained from the former evaluations methods and the considerations made after analysing the responses from crowdsourcing.

### 7.2. AUTOMATIC EVALUATION

The computation of the binary "saliency labels", included in the Political Speeches Dataset, allows the utilisation of supervised Machine Learning highlights for the problem of video segments classification. To this extent, the models attempted in this research are trained in order to minimise the prediction error made in the classification. The performance of the algorithms can be measured in terms of precision and recall, as explained in Section 3.4.1. These values are calculated for the two video segments classes: non salient segments (0) and salient segments (1). However, the dataset is heavily unbalanced, as the amount of video segments labelled 1 is only about 7% of the total. In fact, the political speeches are likely to last longer than one hour, while the news highlights, from which the labels are extracted, have average duration of only three minutes. The data unbalance introduces an additional difficulty in the training process. Indeed, because of that, it is easy to obtain high accuracy, up to 93%, by simply classifying all the video segments as non relevant (0).

Method: *Random Forest*

Metric	Apparent error/accuracy		True error/accuracy	
	Class 0	Class 1	Class 0	Class 1
Accuracy	0.7672		0.6235	
Precision	0.7764	0.7586	0.9312	0.0762
Recall	0.7506	0.7838	0.6419	0.3839
F-Score	0.7633	0.7710	0.7600	0.1272

Table 7.1: Accuracy obtained employing the Random Forest algorithm, trained including the 130 most discriminating features and 100 trees.

Method: *Memory Fusion Network*

Model	Metric	Apparent error/accuracy		True error/accuracy	
		Class 0	Class 1	Class 0	Class 1
<i>mfn_aug_bs64</i>	Mean Absolute Error	0.0023		0.1159	
	Precision	0.9992	0.9973	0.9386	0.0753
	Recall	0.9969	0.9993	0.9386	0.0753
	F-score	0.9981	0.9983	0.9386	0.0753

Table 7.2: Results of the best performing MFN model, trained on the augmented dataset

In spite of the attempts to balance the dataset by using data augmentation, or considering a smaller fraction of the dataset with equal percentage of data points from both classes, the results in terms of precision/recall are quite low for class 1. This problem is shared by all the algorithms experimented in this research, from traditional Machine Learning algorithm to the MFNs. The inability of the models to correctly classify relevant video segments, besides the lack of positive samples, might be due to the fact that the relations among relevant video segments are too abstract to be recognised and modelled by the former algorithms. Moreover, the relations that link the highlight clips collected in the Political Speeches Dataset might not be consistent. For example, when analysing the speeches of the same politician, it might be possible to recognize the usage of certain rhetorical strategies corresponding to the moment in which they are saying something important, as well as certain gesture or tone of the voice. In this research, the speeches from six politicians were included in the dataset. Perhaps, the inclusion of more politicians has introduced an additional obstacle to the creation of a model that is capable to generalise over more speeches, by making the relations that link important statements from different speeches more subtle. In addition to this, the filmmaking style adopted by the news providers might be so different from one another to make it hard to recognize common characteristics in different highlight clips even by a human.

The accuracy of the models that were deployed for the creation of the final highlight clips, presented in the human evaluations, are given in Tables 7.1, 7.2, 7.3. These tables show the performance of the three automatic approaches created for highlights extraction, namely: the unimodal baseline, based on the summarisation of the speech transcripts; the best performing multimodal Machine Learning models based on features concatenation, that is achieved using Random Forests, and on feature integration, i.e. the MFN. They contain the total the scores obtained for precision, recall and f-score, in addition to the accuracy scores or, in the case of MFN, the Mean Absolute Error.

The evaluation in terms of precision and recall does not really apply to the case of the baseline method, based on automatic text summarisation of the speech transcripts. In fact, even though it can be useful to know how many scenes from the ground truth highlights appear on the baseline highlight clips, high overlapping between the two cannot be expected, since the two highlight extraction types are based on different approaches. The baseline method is unsupervised and uses the TextRank summarisation algorithm [131] to seek for central sentences, that have many references across the speech analysed. These sentences are selected and combined to compose the transcript summary, out of which the highlight clip is made. The criteria adopted in the baseline approach might not be the same as the ones used by the professional news reporter when creating the highlight clips. This justifies potentially low precision/recall scores obtained by the baseline, which, for the same logic above, do not pose a problem.

Method: <i>Baseline</i>		
Metric	Class 0	Class 1
Accuracy	0.9026	
Precision	0.9305	0.0968
Recall	0.9675	0.0459
F-Score	0.9486	0.0623

Table 7.3: Scores obtained by the baseline method, based on the transcript summarisation

Despite of the little achievement in the capacity of the employed Machine Learning models to classify the speeches video segments with sufficient precision and recall, the human evaluations can potentially still show promising result. In fact, the baseline method, based on text summarisation, might result as a good unsupervised approach for the identification of salient speech segments. On the other hand, the two chosen multimodal Machine Learning algorithm, namely Random Forest and MFN, might have been trained in order to classify as salient the video segments that present certain characteristics, e.g. the presence of key words, of a certain speech tone or facial expression. Although these features might not be able to represent all the salient scenes from the ground truth highlights, they might be a useful criteria for the creation of highlight clips that somehow result informative, engaging or expressive to the viewers. The results from the human evaluation follow in the next sections.

### 7.3. EVALUATION WITH CROWDSOURCING

Human evaluation provided the ultimate results of this research. Human evaluation was used to measure the general appreciation of the highlight clips generated from the Political Speeches Dataset, using under different criteria, and was implemented through crowdsourcing on the platform Figure Eight [134]. The crowdsourcing tasks design was decided through the process described in the previous Chapter 6, which resulted in two separate surveys, one to assess each generated highlight clip individually and the other one consisting in a pairwise comparison between two highlight clips from the same speech but generated through different methods.

For the sake of budget control, only highlights from 5 different speeches, all belonging to the test set, were considered. Namely:

- 26: Barack Obama - "Address at an Associated Press Luncheon"
- 23: Barack Obama - "Afghanistan Troop Reduction Address to the Nation"
- 82: Donald Trump - "Donald Trump Commissions the USS Gerald R Ford in Norfolk Virginia"
- 61: Donald Trump - "Donald Trump Holds a Political Rally in Pensacola Florida"
- 46: Barack Obama - "Memorial Address for Nelson Mandela"

These videos were chosen in order to include speeches for more than one politician and because of the similar duration of their highlight clips, that have mean of 2.23 minutes and standard deviation of 1.00. All the highlight clips involved in the evaluation are available on YouTube, see Appendix B.1.

According to the analysis conducted in Chapter 6, each task from the two crowdsourcing jobs were completed by six different workers. The workers who took part in the surveys were constrained to belong to Level 3, namely "Highest Quality: Smallest group of most experienced, highest accuracy contributors", as detailed in [134] and to reside in English speaking countries (Australia, Canada, Ireland, New Zealand, United Kingdom and United States). To favour the diversity of the people involved in the crowdsourcing jobs, each worker was allowed to complete only a maximum of 10 tasks.

The jobs corresponding to the two surveys consist in 120 tasks for the individual assessment and 180 for the pairwise comparison. The surveys were published on the 30th of August 2019 and were completed during the same day. The full results from the two jobs and their analysis follow below.

#### 7.3.1. INDIVIDUAL ASSESSMENT

The individual assessment survey consists in two five-level Likert scale questions and two Yes/No control questions, to spot whether the evaluation of the highlight clips is influenced by prior knowledge of the speech and fondness or aversion towards the main speaker. The questions are:

- Q1** *This highlight clip gives me information about what the full speech was about.* How much do you agree with this statement? (-2: *Strongly disagree*, -1: *Disagree*, 0: *Neither agree nor disagree*, 1: *Agree*, 2: *Strongly agree*)
- Q2** *This highlight clip makes me want to watch the full speech.* How much do you agree with this statement? (-2: *Strongly disagree*, -1: *Disagree*, 0: *Neither agree nor disagree*, 1: *Agree*, 2: *Strongly agree*)
- Q3** I like the politician in the video. (*Yes/No*)
- Q4** I have already heard this speech. (*Yes/No*)

All the responses obtained from the Likert-scale questions can be observed in the Tables 7.4 for Q1 and 7.5 for Q2, in which every row corresponds to a job task. Each cell contains the percentage of workers that chose the corresponding vote, calculated out of the six workers who completed the task. These results are reported in order to show that the workers responses vary significantly from one speech to another. For example, in Table 7.4, it can be noticed that for speech 26 at least 50% of the workers agreed on the fact that the four clips are able to convey a sufficient level of information, while speech 46 only the baseline highlight clip was considered informative. These differences can be due to the great diversity of the speeches content, as for some of them the key message might be more explicit and easier to identify without background knowledge. In addition, some of these speeches are very well known by audiences. From the tables, it can also be noticed that for both Q1 and Q2 it is hard to express with certainty which highlights extraction method is better in terms of informativeness and ability to spark interest, according to the workers opinion.

Q1: "This highlight clip gives me information about what the full speech was about."  
How much do you agree with this statement?

Speech	Method	<i>Strongly disagree</i>	<i>Disagree</i>	<i>Neither agree nor disagree</i>	<i>Agree</i>	<i>Strongly agree</i>
26	Ground truth	0.00%	0.33%	0.00%	0.50%	0.17%
	Baseline	0.00%	0.33%	0.17%	0.33%	0.17%
	Random Forest	0.00%	0.00%	0.00%	0.50%	0.50%
	MFN	0.00%	0.17%	0.00%	0.50%	0.33%
23	Ground truth	0.00%	0.33%	0.17%	0.33%	0.17%
	Baseline	0.00%	0.50%	0.17%	0.17%	0.17%
	Random Forest	0.00%	0.33%	0.00%	0.17%	0.50%
	MFN	0.00%	0.17%	0.00%	0.67%	0.17%
82	Ground truth	0.00%	0.33%	0.33%	0.17%	0.17%
	Baseline	0.00%	0.17%	0.17%	0.50%	0.17%
	Random Forest	0.00%	0.50%	0.17%	0.33%	0.00%
	MFN	0.00%	0.50%	0.33%	0.00%	0.17%
61	Ground truth	0.17%	0.33%	0.00%	0.33%	0.17%
	Baseline	0.00%	0.50%	0.00%	0.33%	0.17%
	Random Forest	0.00%	0.50%	0.00%	0.17%	0.33%
	MFN	0.00%	0.17%	0.17%	0.33%	0.33%
46	Ground truth	0.00%	0.67%	0.17%	0.00%	0.17%
	Baseline	0.00%	0.33%	0.00%	0.17%	0.50%
	Random Forest	0.00%	0.67%	0.00%	0.00%	0.33%
	MFN	0.00%	0.83%	0.17%	0.00%	0.00%

Table 7.4: Responses to the first Likert question for all the speeches.

The analysis of the answers from the Likert-scale questions was conducted following the recommendation from [143]. First, the results from Q1 and Q2 were interpreted by converting the 5 levels of the Likert scale (*Strongly disagree*, *Disagree*, *Neither agree nor disagree*, *Agree*, *Strongly agree*) into their corresponding numeric values (-2, -1, 0, 1, 2) and calculating the mean response. This results in the bar graphs in Figures 7.1a and 7.1b. Even here it is noticeable how the results are different from one speech to another. In fact, regarding Q1, the highlight clips generated for speeches 26, 23 and 61 all have neutral or positive mean responses, while for speeches 82 and 46 only the baseline highlight clips are able to convey a sufficient amount of information, according to the crowdworkers.

Q2: "This highlight clip makes me want to watch the full speech (independently on my political preference or my interest in politics)."  
How much do you agree with this statement?

Speech	Method	Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
26	Ground truth	0.00%	0.00%	0.67%	0.17%	0.17%
	Baseline	0.00%	0.17%	0.33%	0.50%	0.00%
	Random Forest	0.00%	0.17%	0.17%	0.33%	0.33%
	MFN	0.00%	0.00%	0.50%	0.33%	0.17%
23	Ground truth	0.00%	0.33%	0.17%	0.33%	0.17%
	Baseline	0.00%	0.00%	0.67%	0.33%	0.00%
	Random Forest	0.00%	0.00%	0.33%	0.17%	0.50%
	MFN	0.17%	0.00%	0.33%	0.17%	0.33%
82	Ground truth	0.00%	0.17%	0.17%	0.50%	0.17%
	Baseline	0.17%	0.17%	0.17%	0.33%	0.17%
	Random Forest	0.00%	0.17%	0.50%	0.33%	0.00%
	MFN	0.00%	0.00%	0.67%	0.33%	0.00%
61	Ground truth	0.17%	0.33%	0.33%	0.17%	0.00%
	Baseline	0.00%	0.17%	0.67%	0.00%	0.17%
	Random Forest	0.17%	0.33%	0.17%	0.00%	0.33%
	MFN	0.00%	0.33%	0.00%	0.17%	0.50%
46	Ground truth	0.00%	0.00%	0.50%	0.33%	0.17%
	Baseline	0.00%	0.00%	0.33%	0.33%	0.33%
	Random Forest	0.00%	0.00%	0.50%	0.50%	0.00%
	MFN	0.00%	0.17%	0.50%	0.33%	0.00%

Table 7.5: Responses to the second Likert question for all the speeches.

It can be argued that the baseline summaries are never considered non informative, as the baseline method, which is employed for their generation, is based on the transcript summarisation. Given the validity of TextRank algorithm [131] used for the summarisation, highlight clips based on the speech summarisation are likely to contain numerous references to the main issue outlined in the speech. For this reason, baseline highlight clips might be considered more informative than the real highlights extracted from the news, that here are missing the further explanations and discussion that are often given by the newsreader, before or after the broadcast of the highlight clips. What can also be noticed at first glance from plot Figure 7.1a is that for every speech at least one type of automatically generated highlight clips have higher mean score than the ground truth highlights, which is promising for the perceived quality of the automatic highlights. For what concerns the mean responses to Q2, visible in Figure 7.1b, it can be noticed that, for most of the highlight clips, the ability to arouse interest in the viewer to watch the full-length speech is positive and never negative. The lowest scores are obtained by the highlight clips of speech 61, but this is likely due to the nature of this particular speech by Donald Trump.

In order to achieve a overall vision of the relative informativeness and ability to spark interest of the different highlight clips, the responses for the different speeches were aggregated by counting the number of times each vote was assigned for the highlight clips belonging to the same category. The frequency of the votes is expressed in percentage and the results are visible in the pie chart in Figure 7.2. For Q1, it is noticeable that, to almost a parity of disagreement, the highlight clips generated from the Random Forest algorithm are the ones that obtained the highest amount of positive votes, while the lowest amount corresponds to the ground truth highlights. This positive result shows that highlight clips extracted from Machine Learning techniques are able to outperform ground truth highlights in terms of amount of relevant information conveyed. Regarding Q2, it is noticeable that a big fraction of crowdworkers did not express a clear opinion (*Neither agree or disagree*) about whether the clips sparked in them interest in watching the full video. However, the pie charts in Figure 7.2b show that at least 43% of the workers involved in these tasks would like to watch the full speech. The answers to this question might be influenced by the politician reputation, by the fact that they might have heard the speech already, or by the nature of the speech itself, that might be considered boring or unimportant.

Since it is hard to compare the the highlights extraction methods from the pie charts in Figure 7.2, the responses were further aggregated by summing all the votes, that range from -2 (*Strongly disagree*) to +2 (*Strongly disagree*), obtained for every highlight clips type. The final results are shown in Figure 7.3. From the two pie charts it is possible to obtain a ranking of the four highlights extraction methods. In terms of informativeness, Figure 7.3a, the methods based on Random Forest is on the podium, followed by baseline highlights made from transcript summarisation, highlights generated utilising MFNs, and finally ground truth highlights. In terms of ability to spark interest, the first method is still Random Forest, followed by MFN and baseline highlight clips and, lastly, again ground truth highlights. The classification confirms that it is possible to automatically generate highlight clips that are at least as informative and as able to create interest as highlight clips replicated from the ones broadcast by real news channels. This represents a positive answer to **RQ2**, not only because Machine Learning was effectively applied for automatic highlights extraction from videos where one person speaks facing the camera, but also because multimodal methods resulted in more informative highlight clips than the unimodal baseline.

Finally, the Pearson correlation coefficients  $r$  was calculated from the responses to the Likert questions and the two control questions, to check whether the answers to the four questions are somehow correlated. The Pearson correlation coefficient is defined according to the formula<sup>1</sup>:

$$r = \frac{\sum(x - m_x)(y - m_y)}{\sqrt{\sum(x - m_x)^2(y - m_y)^2}},$$

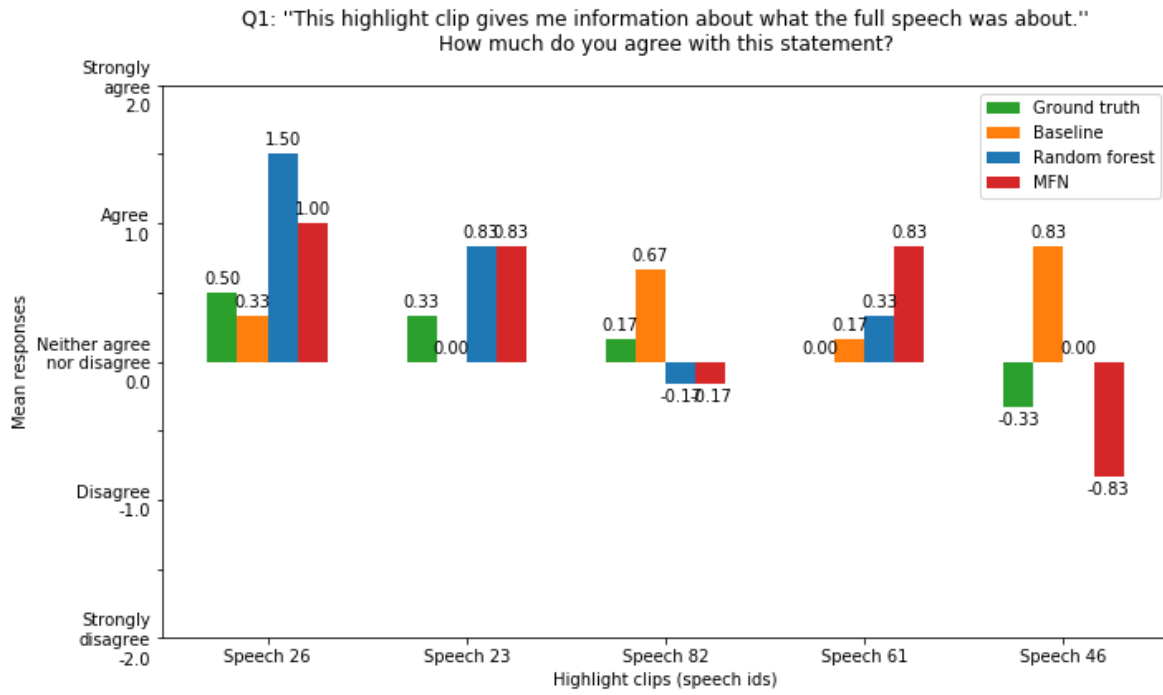
where  $x$  and  $y$  are the two votes distributions and  $m_x$  and  $m_y$  is their mean. The results are reported in Table 7.6, that shows both the correlation coefficients and the p-values. What is interesting to notice is that the answers to the Likert questions are weakly positively correlated (0.56), as well as Q2 with Q3 ("I like the politician in the video." (*Yes/No*)). In fact, it is reasonable to think that a user would be more encouraged to watch the video of the full-length speech of a politician that she likes, rather than dislikes. On the other hand, there is a weak negative correlation between the responses to Q1 and Q4 ("I have already heard this speech." (*Yes/No*)), which can be motivated by the fact that if a viewer has background knowledge about the speech, she might be more inclined to notice all the information that the highlight clip does not cover.

In addition to this, the correlation coefficient between the responses to the Likert questions and the duration of the highlight clips involved in the human evaluation was calculated. However, no significant correlations were discovered, as can be seen from Table 7.7. This represents a successful outcome, as it means that the duration of the highlight clips was not a cause of bias in the responses. For example, longer highlight clips do not necessarily increase the amount of relevant information conveyed.

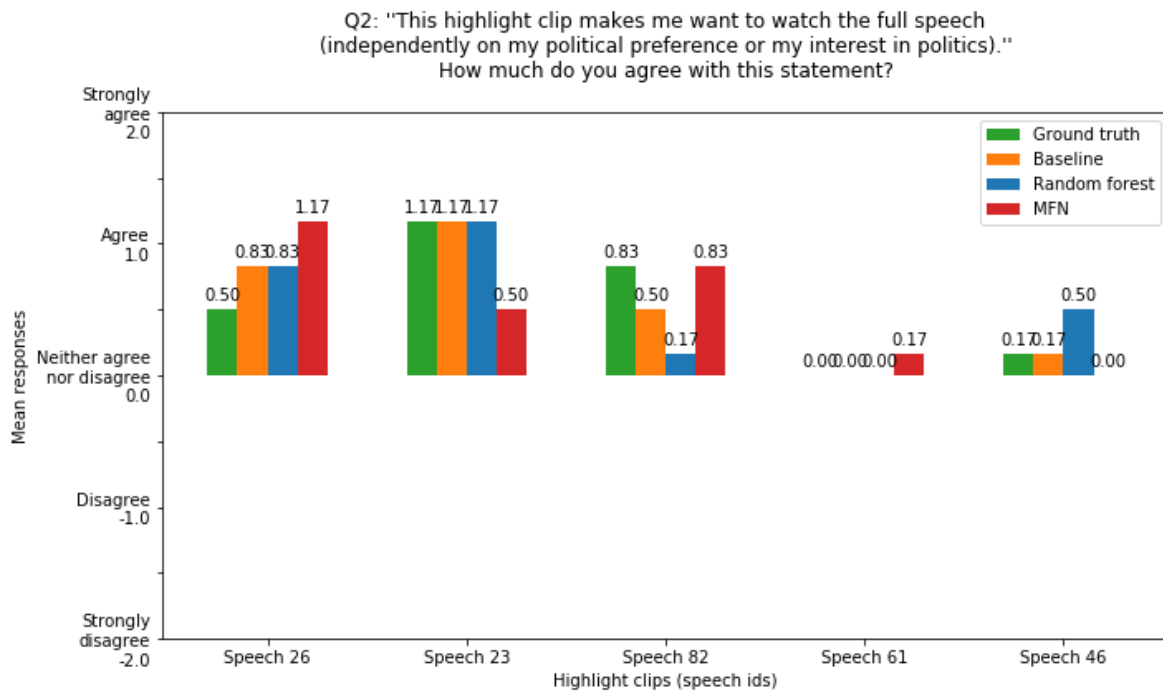
The results concerning additional criteria of evaluation of the highlight clips are provided in the analysis of the responses to the pairwise comparison crowdsourcing tasks.

<sup>1</sup><https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.pearsonr.html>





(a) Q1



(b) Q2

Figure 7.1: Mean responses for speech clip and video type

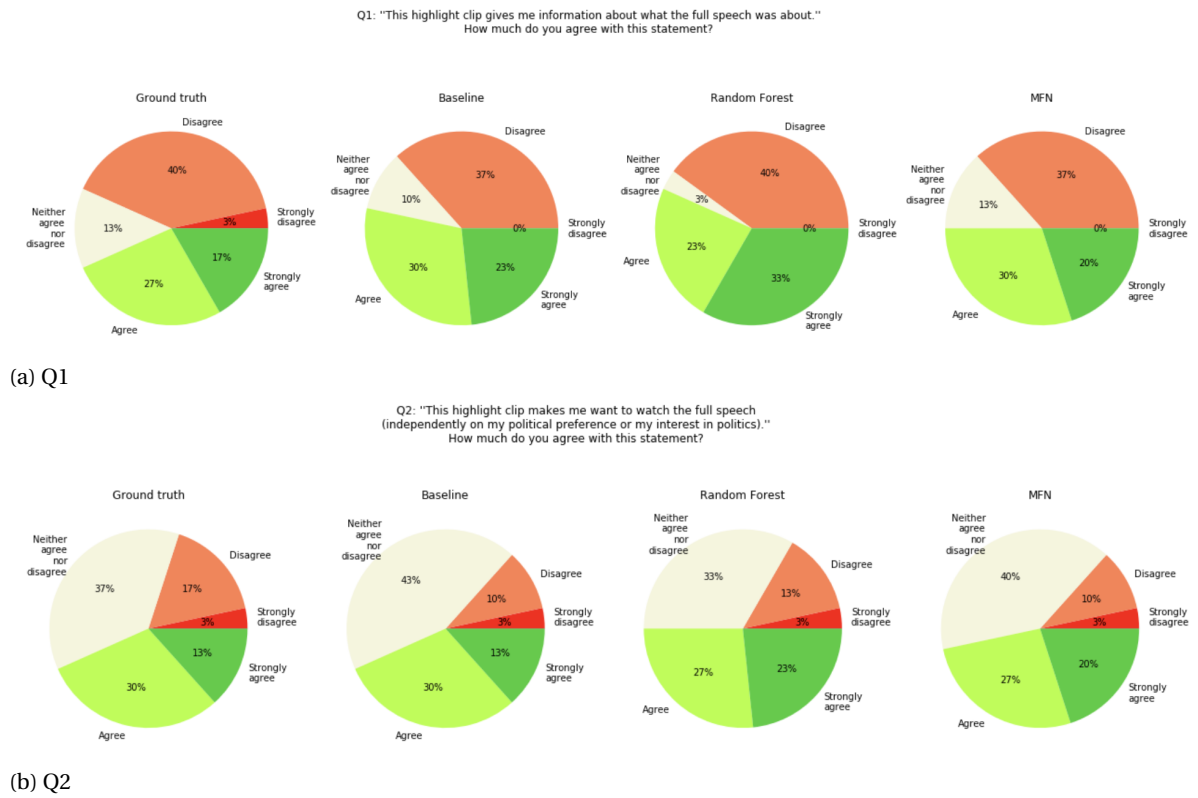


Figure 7.2: Aggregated responses to the Likert questions for the four highlights extraction methods

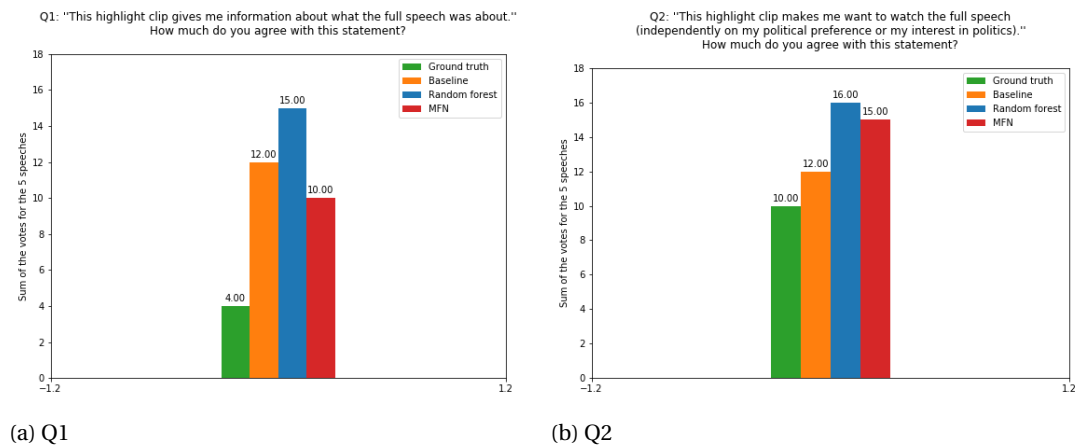


Figure 7.3: Final aggregation of the responses obtained for the two Likert questions

Questions	Pearson's correlation coefficient	Two-tailed p-value
Q1 - Q2	0.56	0.00
Q1 - Q3	0.05	0.62
Q1 - Q4	-0.31	0.00
Q2 - Q3	0.32	0.00
Q2 - Q4	-0.08	0.39

Table 7.6: Correlation between the responses of the questions from the individual assessment

Questions	Pearson's correlation coefficient	Two-tailed p-value
Q1	0.08	0.75
Q2	-0.08	0.74

Table 7.7: Correlation between the responses to the Likert questions from the individual assessment and the duration of the highlight clips

### 7.3.2. PAIRWISE COMPARISON

The reason why the pairwise highlight clips comparison was included in the human evaluation is that the responses from the solely individual assessment might not be enough to extract a distinct ranking of the highlights extraction methods. For example, the answers to the Likert questions might result in ties between different highlight clips types. These ties cannot happen in the pairwise comparison tasks, as they are designed in such a way that the worker is always required to select one highlight clip out of two.

The pairwise comparison survey consists in comparing pairs of highlight clips and selecting the one that appears to be a more suitable answer to the questions:

**Q1** Which video gives more information about the topic of the full speech?

**Q2** Which video contains more repetitions of the same sentences?

**Q3** Which video is more engaging (independently on its content or the political tendency)?

**Q4** Which video is more cinematic?

**Q5** In which video is the speaker more expressive?

To be able to give a numerical value to the answers, one point was assigned a highlight clip type every time this was chosen in a comparison. Since Q2 allows the option "*The two clips contain the same amount of repetitions*", when this was selected +0.5 was added to both the highlight clips involved in the comparison. The results for the five questions, divided by speech, is visible in Figures 7.4, 7.5, 7.6, 7.7, 7.8. As occurred in the individual assessment, even in these results it is clear that the responses are very different from speech to speech and there is not a clear trend. What can be seen is that for each speech one particular highlight clip stands out from the rest. That is: ground truth highlights for speech 26 and 46, Random Forest highlights for speech 23, baseline highlights for speech 61 and Random Forest and baseline highlights compete for speech 82. This confirms that, even though there is not a noticeable superiority of the automatic highlights compared to the ground truth, for three out of five speeches the scenes contained in the automatic highlights are considered more informative, engaging and cinematic and the speaker more expressive.

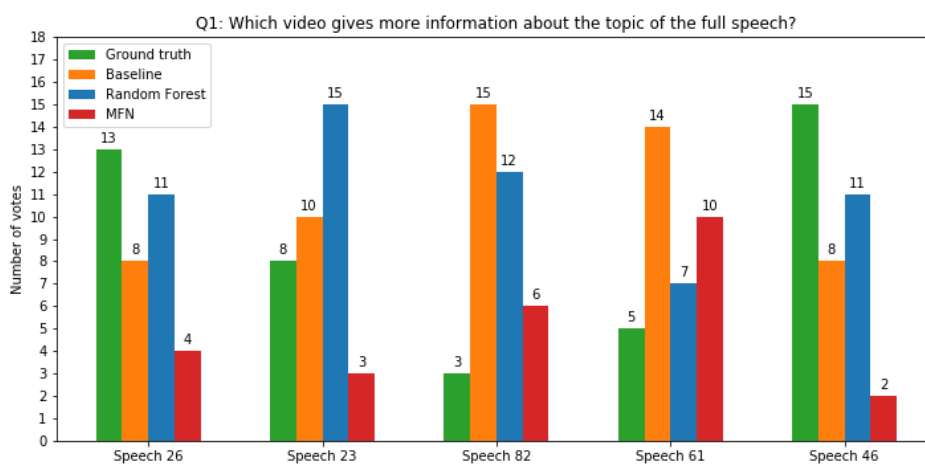


Figure 7.4: Q1

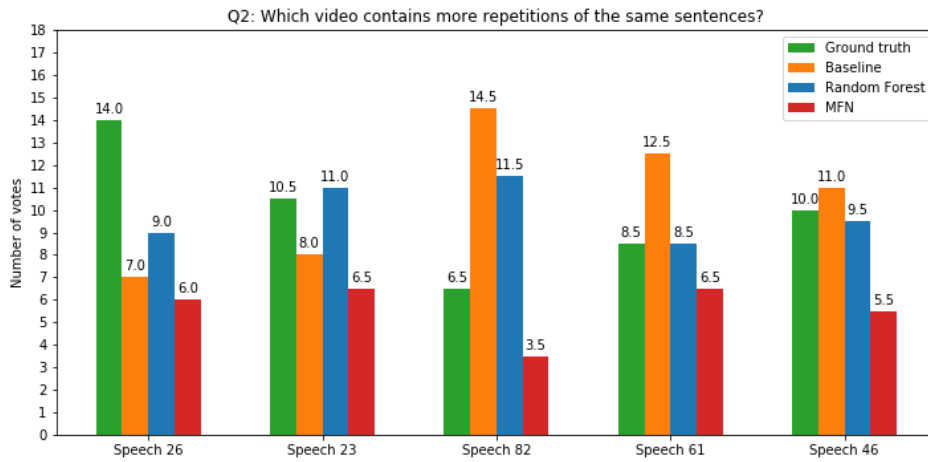


Figure 7.5: Q2

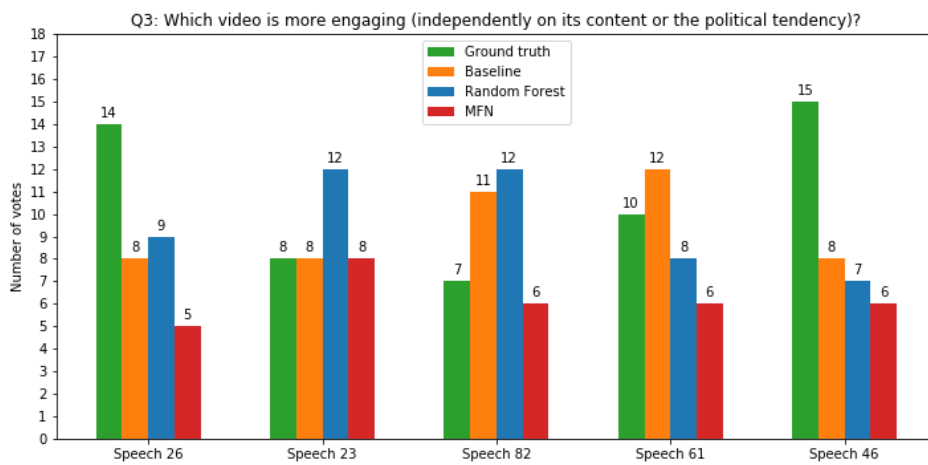


Figure 7.6: Q3

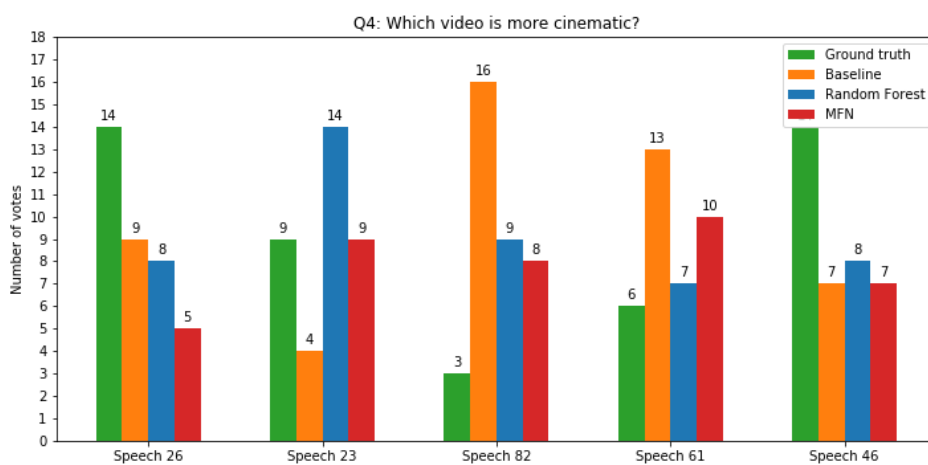


Figure 7.7: Q4

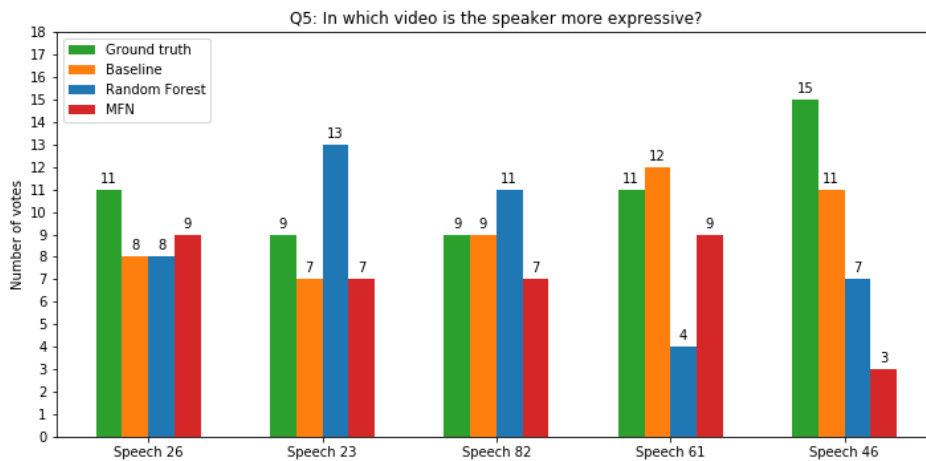


Figure 7.8: Q5

In order to obtain an overall ranking of the four highlight extraction methods, the points for each type were summed together. The visual representation of points expressed in percentage can be found in the pie charts in Figure 7.9. In all the charts, the percentages obtained by the ground truth, baseline and Random Forest are similar, ranging from approximately 24% to 31%. Instead, highlight clips generated with MFN obtained lower scores. As for what concerns Q1, the highlight clips generated with Random Forest still turn out to be the most informative, beating the ground truth highlights. The second scoring method is the baseline. This is a very good outcome, since not only two automatically generated highlight clips are able to overcome ground truth highlights in terms of informativeness, but also a highlight extraction method based on multimodal Machine Learning (Random Forest highlights) beats the baseline, solely based on the text summarisation of the speech transcripts.

Again these results verify the validity of the research hypothesis: **Hp**: *Multimodal features can be used to train Machine Learning models that give better results in information extraction from videos, compared against unimodal methods..* In fact, it was proven that that the most informative highlight clips created in this research are the ones produced utilising a multimodal approach based on the Random Forest algorithm, which beats both the ground truth highlight clips and the unimodal baseline highlight clips. This achievement was validated both from the individual assessment survey and the pairwise comparison survey.

Moreover, the results of Q1 confirm the necessity of including a pairwise comparison of the generated video clips in the evaluations. In fact, if MFN highlight clips came in third in the individual assessment, in the pairwise comparison it was only considered "more informative" the 14% of times. This means that the MFN highlight clips themselves have a sufficient level of informativeness (see Figure 7.2a), yet they are not comparable to the other methods.

Regarding Q2, the most repetitive highlights extraction method results to be the baseline. This can be motivated by the fact that the TextRank algorithm [131], on which the text summarisation is based, looks for key words or key sentences in a text and selects the parts of the text that are linked to these. A connection between sentences exists if they have a certain level of similarity. The similarity function is implemented to measure to what extent two sentences overlap. Because of this, it can be understood that the sentences that compose the transcript summaries contain repetitions of the key words or sentences identified by TextRank. In spite of this, the baseline highlight clips also resulted to be the most cinematic among the experimented methods and scored only 4% points less than the ground truth in terms of viewers engagement. Finally, ground truth highlights still achieved more consensus than the others in terms of engagement and expressiveness of the main speaker, even though their percentages were close to the ones obtained by the baseline and Random Forest highlights clips.

Similarly to the individual assessment, the Pearson correlation coefficients  $r$  between the responses to the questions was calculated. The results are reported in Table 7.8. All the questions couples are somehow weakly positively correlated. The most significant correlations, where  $r$  is 0.75 or greater, can be found for the pairs Q1-Q2, Q1-Q3, Q1-Q4 and Q2-Q3. This shows that the highlight clips that convey a greater amount of information also seem to contain more repetitions, to be more engaging and cinematic. This might be interpreted as a highlight clip that contains relevant information is naturally perceived as more engaging

and also cinematic. But also, more information can imply more sentences included in the highlight clips and, therefore, the result might be considered a bit repetitive. The weak correlation between the amount of repetition and the level of engage is not meaningful.

Finally, the correlations between the duration of the highlight clips and the responses to the five questions was calculated but without finding any significant trends. The results are reported in Table 7.9.

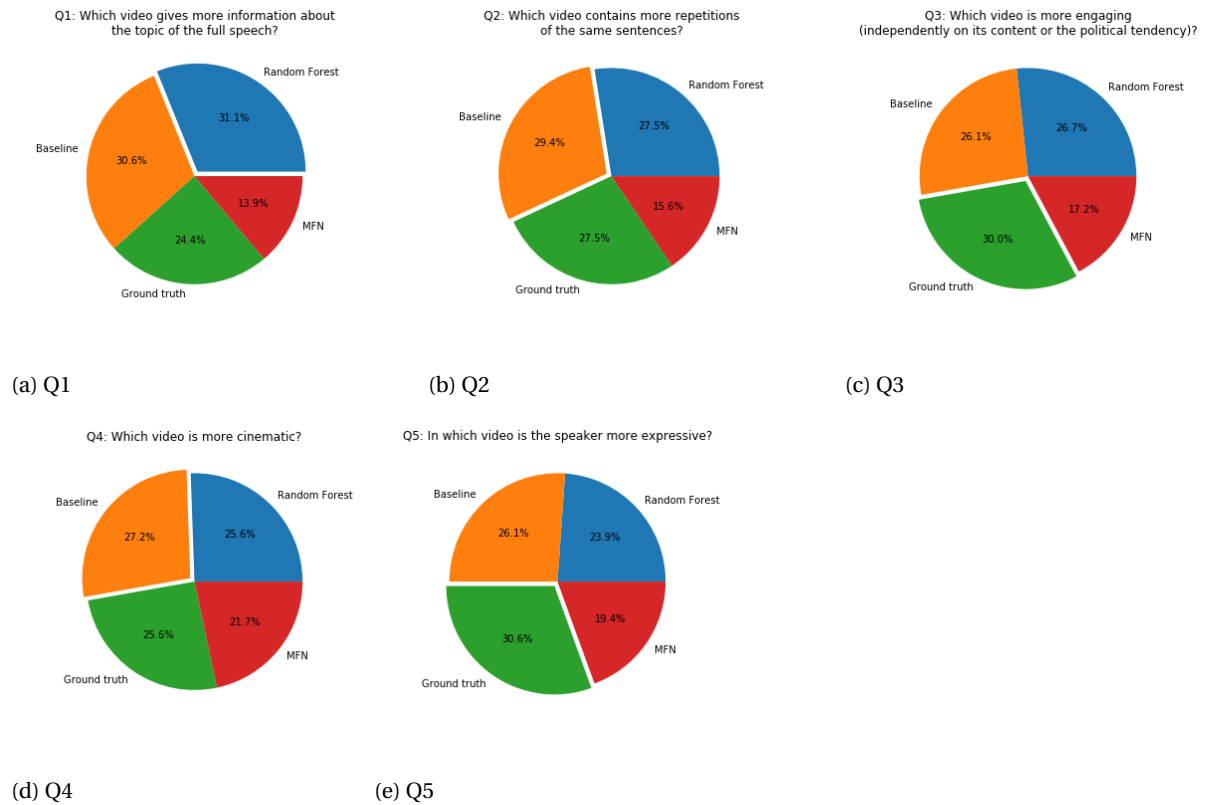


Figure 7.9: Aggregated responses to the questions Q1, Q2, Q3, Q4 and Q5

Questions	Pearson's correlation coefficient	Two-tailed p-value
<b>Q1 - Q2</b>	0.77	0.00
<b>Q1 - Q3</b>	0.75	0.00
<b>Q1 - Q4</b>	0.78	0.00
<b>Q1 - Q5</b>	0.61	0.00
<b>Q2 - Q3</b>	0.76	0.00
<b>Q2 - Q4</b>	0.67	0.00
<b>Q2 - Q5</b>	0.54	0.01
<b>Q3 - Q4</b>	0.72	0.00
<b>Q3 - Q5</b>	0.73	0.00
<b>Q4 - Q5</b>	0.51	0.02

Table 7.8: Correlation between the responses to the questions from the pairwise comparison

Questions	Pearson's correlation coefficient	Two-tailed p-value
Q1	0.01	0.97
Q2	0.05	0.83
Q3	-0.01	0.95
Q4	-0.13	0.59
Q5	0.17	0.48

Table 7.9: Correlation between the responses to the questions from the pairwise comparison and the duration of the highlight clips

## 7.4. DISCUSSION

This chapter presented the ultimate results obtained in the present research about automatic highlights extraction from videos. The experimentation with different Machine Learning methods for highlight extraction gave origin to three types of automatically generated highlight clips for each speech contained in the Political Speeches Dataset. These methods include a baseline unimodal approach, based on the transcript summarisation, compared against two multimodal Machine Learning models, that are trained using a combination of textual, audio and visual features. In these two algorithms, namely Random Forest and MFN, two different ways of multimodal features integration are experimented. In the Random Forest algorithm, the three multimodal features category are simply joined by concatenating the respective matrices on one dimension. On the other hand, MFN employs a more sophisticated approach, in which the features are first processed separately by three distinct LSTM neural networks, from which multimodal hidden dynamics are extracted by seeking for interactions between the memories of the different LSTM cells over time.

If the baseline relies on an unsupervised algorithms, namely TextRank [131], which builds summaries from a set of central sentences selected from the speeches transcripts, the other two methods are trained in order to identify the speeches video segments that appear in the ground truth highlight clips. The training is performed exploiting the binary "saliency labels" available in the Political Speeches Dataset, by trying to predict the segments labels committing the minimum misclassification error. The performance of the algorithms can be measured in terms of precision and recall. Tables 7.1 and 7.2 show the results obtained by the two chosen best performing models. Despite the attempts to improve the accuracy through data augmentation, by adding penalties to the loss function for the misclassification of segments from class 1 (salient segments), and dropout and early stopping in the case of MFN models, it was not possible to obtain a sufficient accuracy in the prediction of the salient segments. This is likely to be due to the great difference between the speeches and the highlight clips that compose the Political Speeches Dataset: not only the politicians involved have different rhetorical strategies and debate about a wide range of topics, but also the highlight clips by different news channels are characterised by their particular filmmaking style, not to mention the political bias which might affect the perception of which speech segments to consider significant.

All these aspects imply that the answer to **RQ1**, considering the case of the videos from the Political Speeches dataset, is negative. In fact, the statement that the videos of the speeches share a common structure, easy to identify and analyse, is probably not true. Therefore, the design a statistical model which describes the distribution of the salient segments is not possible to obtain. Collecting a wider range of labelled video segments from political speeches might be a solution for achieving better performances. Regarding the MFN architecture, the accuracy might be improved by introducing additional layers in the network, that can further process the feature matrix and extract higher level representations. Currently, the Random Forest algorithm that uses simple concatenation to join multimodal features is the best option. This answers the second research subquestion of **RQ2**, that is "*What is the most effective method for the integration of multimodal features?*"

Even though the utilised Machine Learning algorithm failed in classifying the video segments according to the ground truth, the results achieved from the human evaluation, implemented through two crowdsourcing jobs, shows promising results. First, the baseline highlight clip based on the transcripts summarisation are considered more informative, cinematic and able to spark interest in the viewers (see Figures 7.3a, 7.9d, 7.3b). This means that an unsupervised method for the creation of good quality highlight clips, or at least comparable to the quality of the ground truth, was discovered. In addition to this, the first multimodal algorithm, the Random Forest, outperformed the baseline in terms or informativeness and ability to generate interest (see Figures 7.3a, 7.3b, 7.9a). This last result represents a great achievement, because it proves that the combination of more feature modalities can lead to better performances than unimodal methods, like the ones only based on the speech content, and represents a positive answer to the first research subquestion of **RQ2**, that is "*When using Machine Learning for video analysis with the purpose of information extraction, do*

*the combination of textual, audio and visual features outperform purely content-based methods?". This also confirms the validity of the main research hypothesis: **Hp**: *Multimodal features can be used to train Machine Learning models that give better results in information extraction from videos, compared against unimodal methods..**

The fact that the highlights automatically extracted with multimodal Machine Learning obtained consent from the workers, although they do not resemble the ground truth, can be explained considering that the Random Forest and MFN models might learn to predict based on certain properties of the speech tone and the facial expressions used, or the occurrence of certain words or sentences, that are not sufficient criteria to obtain high precision/recall score, but might be suitable for the identification of video segments that are worth including in the highlight clips.

Finally, the calculation of the correlations among the responses to the crowdsourcing tasks and the highlight clips duration demonstrates the absence of relevant biases that affected the human evaluation. In particular, the analysis of the two control questions from the individual assessment, Q3 and Q4, has shown that background knowledge on the speech and political tendency of the workers did not cause a bias in the evaluation. Similarly, the duration of the clips does not imply greater informativeness. Therefore, it can be concluded that the design of the crowdsourcing tasks and the choice of the involved highlight clips made it possible to obtain fair evaluations.

This last statement provides the answers to **RQ3** (*Is crowdsourcing an effective method for the evaluation of the results of automatic information extraction from videos?*). Moreover, regarding the respective research subquestion (*If crowdsourcing can be used for the evaluation of the results, how should the tasks included in the crowdsourcing process be designed?*) it can be confirmed that the tasks designed for the individual assessment and pairwise comparison surveys compose a valid framework for the evaluation of the results of automatic information extraction from videos.

For this reason, an analogous human evaluation can be repropose for future research about the topic of automatic highlights extraction from videos.



# 8

## CONCLUSION AND FUTURE WORK

### 8.1. CONSIDERATIONS AND CONCLUSIONS

In this thesis, the problem of automatic information extraction from videos has been introduced and explored. Whereas other areas of research in Machine Learning have been extensively investigated, leading to the achievement of outstanding results in information extraction from different sources, as in the case of text analysis with NLP or object detection in images, there still remains plenty of room for exploration within the case of video understanding.

Considering the broadness and complexity of the latter problem, the technologies and the amount of labelled data that are available nowadays are still relatively limited. These restrictions represent a major obstacle for the creation of a sole algorithm that can generalise over videos of different nature and concerning different topics. In order to overcome such difficulties, in several research projects — including the ones that were presented in Chapter 2 and that have inspired the methodology of this thesis (Chapter 3) — the process of information extraction was simplified by focusing on particular kinds of videos and precise lower level objectives. For example, several attempts of video summarisation, as in [50, 53, 130], focused on the identification of key frames, that represent the information conveyed by the analysed video segments. Another case is given by [11], where the authors delved into the problem of "tropes" recognition in films, especially for the particular kind of horror movies, where these patterns are more explicit and, therefore, easy to distinguish.

Similarly, the challenge of information extraction from videos was tackled in this thesis from the particular perspective of highlights extraction. In order to place this scientific research in the context of the Avengers Project, conducted at IBM Center for Advanced Studies, the research question and methodology of the thesis were designed in such a way to adhere to the process pipeline introduced in Chapter 1 and visible in Figure 1.1. This aspect motivates the choice of dealing precisely with videos in which one person speaks in front of a camera. Automating the process of analysis of this specific kind of videos is important in the industrial context because it can be harnessed for several interesting applications, such as the automatic video summarisation of interviews or the automatic creation of personal video curricula vitae.

In spite of the availability of the MOSI and MOSEI datasets [37, 38], that consist in collection of videos of the form which is in the interest of this research, the lack of labels that indicate the extent of relevance of distinct video segments, in the context of the full-length video, led to the collection of a novel dataset, namely the Political Speeches Dataset. The data collection and labelling process was partially automated, thus allowing its realisation in the restricted time available. The Political Speeches Dataset is unique of its kind. In fact, it is the first dataset containing videos where one person speaks in front of the camera and provided with "saliency labels". The new dataset makes it possible to directly train supervised Machine Learning algorithms for the task of relevant information detection in video segments.

The introduction of a novel dataset represents the greatest achievement obtained in this thesis project. However, additional contributions were made. First of all, two former researches on multimodal learning applied to videos, namely [11] and [12], were replicated, adapting them to the case of video segments classification based on saliency. Therefore, it can be concluded that the methods and techniques introduced in these previous works can be transferred to other prediction problems. Furthermore, one unimodal Machine Learning approach, solely based on transcript summarisation, was compared against two multimodal methods, under the hypothesis that multimodal features constitute a richer representation, that can benefit to the

models classification performance. Even though it was not possible to observe a substantial improvement in the results of multimodal methods, rather than unimodal, it can be argued that the techniques adopted in this first attempt, as well as the Political Speeches Dataset, can be further improved and extended in future work. Nevertheless, the methodology that was presented in this thesis can be repurposed, as the human evaluations proved it to be effective.

Ultimately, in this dissertation it was pointed out that one of the greatest challenges to cope, when dealing with problems like automatic information extraction from videos or video summarisation, is the lack of an effective evaluation method. The final contribution brought by this research is the design and employment of two crowdsourcing processes, implemented using the platform [134], which led, to the extent possible, to a thorough and unbiased evaluation of the highlight clips generated in this work.

## 8.2. ANSWERS TO THE RESEARCH QUESTIONS

Given an overview of the objectives and challenges addressed in this thesis, it is now possible to explain the thesis contributions in relation to the research questions defined in Chapter 1. The set of limitations and problems that came to light while answering to these questions are discussed in the next sections, which offers recommendations for future research.

**RQ1** *Do the salient moments of a video where one person speaks facing the camera share common properties, i.e. the recurrence of particular images, actions, sounds or verbal expressions, that can be identified, classified and used to categorize the scenes of the video?*

Previous research confirmed that the hypothesis that videos belonging to the same category are characterised by common properties is true to some extent. For example, in [11], the authors proved the existence of "tropes", namely storytelling strategies that suggest to the spectators what is going to happen, which can be exploited to classify the scenes of horror movies according to their content. This is utilised by professional filmmakers for the composition of movie trailers.

However, in the case of videos where one person speaks facing the camera it can be concluded that common patterns do not exist, or, if they do, they are too subtle or abstracted to be identified with certainty. Considering the movie reviews included in the MOSI dataset [38], most of the times the speakers follow their stream of consciousness, and a segmentation and categorisation of these videos becomes difficult. Nevertheless, to what concerns the case of political speeches, it is hard to think that the speeches do not contain an underlying structure. Even though common patterns might exist, some aspects of the complexity of this analysis need to be taken into account. First, the topics treated in political speeches are broad, and some of them might be meant to address an audience of experts of the subject matter. Moreover, different politicians use different linguistic registers and rhetorical strategies. Due to these aspects, it is not possible to identify structures in videos of political speeches that allow to easily recognise where salient moments occur.

- *If they exist and are identifiable, can the salient scenes contained in a video where one person speaks facing the camera be used to represent the relevant information contained in the video? For example, can this be done through the creation of a short highlight clip that shows the most significant moments?*

In case the analysed videos contain salient moments that are identifiable, these can be collected and arranged into a highlight clip. Highlight clips are considered as a valid method to convey information. This was confirmed by former research projects [6, 27], as well as by the human evaluation conducted in this thesis. In fact, the results of the highlights assessment show that, on average, the level of informativeness conveyed by the highlight clips is considered sufficient (see Figures 7.1a and 7.2a) to understand what the political speech in question was about.

**RQ2** *If the research question RQ1 is verified as true, is it possible to automatize the process of relevant information extraction from videos where one person speaks facing the camera using Machine Learning and Deep Learning techniques?*

In this research, three Machine Learning methods for the extraction of highlights from videos were explored. The adoption of these methods yielded to the creation of three types of automatically generated highlight clips.

Even though, due to the fallacy of **RQ1** in the case of the videos of political speeches from the Political Speeches Dataset, it was not possible to identify explicitly what features characterise relevant moments in the videos examined, the results of human evaluation demonstrated that the highlight clips, produced by the methods that were experimented, are to some extent informative, engaging and cinematic, thus comparable to ground truth highlight clips, replicated from professional news media. This confirms that automatic highlights extraction from videos is feasible and that future research can lead to the successful achievements of further improvements.

- *When using Machine Learning for video analysis with the purpose of information extraction, do the combination of textual, audio and visual features outperform purely content-based methods?*

The results of human evaluation show that the most informative highlight clips created in this research are the ones produced utilising a multimodal approach based on the Random Forest algorithm. This achievement is hard to explain, as the algorithm was not able to classify the video scenes with high accuracy. However, it can be assumed that the algorithm is able to classify the video segments based on some hidden patterns in the multimodal features that are effective for the realisation of convincing highlight clips. Despite the fact that the overcoming of the baseline method by a multimodal method is not so evident, the result in terms of informativeness obtained by the Random Forest highlight clips can be seen as an encouragement to further explore multimodal Machine Learning for the purpose of video segments classification.

- *If so, what is the most effective method for the integration of multimodal features?*

One of the hypothesis made in this research was that the performance of multimodal Machine Learning based on the simple concatenation of features from different sources of information can be outperformed by a more sophisticated features combination, as in the case of MFNs. However, both the approaches based on Random Forest and MFNs resulted in scarce classification accuracy. This is probably to the fallacy of **RQ1** in the context of the videos of the Political Speeches Dataset. Because of this, it was not possible to assess the performance of the two multimodal approaches fairly. Nevertheless, the results of human evaluation show that highlight clips generated using the Random Forest algorithm, generally achieve more consent than MFN highlight clips. Thus, it can be concluded that, in this case, multimodal features concatenation is a more suitable technique. This conclusion might change if the MFN model were improved and extended with additional neural layers.

### **RQ3** *How can the results of automatic information extraction from videos be evaluated?*

In this thesis it was shown that the performance of information extraction from videos can be evaluated in two ways. Firstly, if the data analysed include labels about the extent of information conveyed by the video segments, a low level evaluation based on the calculation of precision and recall scores can be applied. Secondly, in addition to the former basic evaluation, a higher level evaluation can be achieved from the involvement of human judgement. This can be obtained from the completion of surveys containing questions designed appropriately. Such human evaluation can be implemented through the use of crowdsourcing. Because of the positive results achieved through the human evaluation conducted in this research, crowdsourcing can be considered an effective method for the evaluation of the results of automatic information extraction from videos. In fact, the efficacy of the tasks involved proved that it is possible to design the crowdsourcing jobs in a way that the tasks are simple enough to be executed by the workers with a low risk of error, but, at the same time, allow to obtain thorough information about the quality of the results.

However, crowdsourcing has limitations that need to be taken into account. Most importantly, crowdsourcing usually involves a monetary compensation for the workers who collaborate in the tasks completion. Budget limitations might constitute an obstacle to the collection of a large enough number of responses.

- *If crowdsourcing can be used for the evaluation of the results, how should the tasks included in the crowdsourcing process for the evaluation of the results be designed?*

An effective crowdsourcing study can be achieved by including tasks where the crowdworkers are required to watch the videos resulting from the research and answering questions about the videos quality. Good level results can be obtained only if the duration of the videos is short

enough to maintain the crowdworker focused until their termination. In addition, the questions should be clear and easy to answer, as in closed-ended questions, that lead to easily interpretable answers. Finally, in chapters 6 and 7, the importance of including two types of crowdsourcing tasks was underlined. The inclusion of both an individual assessment and a pairwise comparison allows to obtain absolute and relative information about the quality of the results, which is fundamental for a thorough understanding of the results.

### 8.3. RECOMMENDATIONS AND SUGGESTIONS FOR FUTURE WORK

In the previous sections several difficulties and limitations discovered in this research were underlined. The presence of these obstacles confirms the necessity to continue the research on automatic information extraction from videos, as there is plenty of space for improvement.

#### 8.3.1. EXPERIMENT WITH MORE STRUCTURED VIDEO TYPES

As was formerly pointed out, it is still not possible to design one sole Machine Learning model that can be applicable to every video category. This is mainly due to the actual limitations of the existing techniques, that fail to achieve a high level understanding of the dynamics occurring in a video. Improving the current state-of-the-art is extremely difficult, also because of the lack of useful training data. The scarcity of large-scale datasets containing videos and useful labels for the video analysis, especially the saliency labels that can be used for highlights extraction, is a major obstacle. In fact, Machine Learning algorithms need a large amount of data in order to be trained effectively. Nonetheless, high quality achievements with a particular video type, as the videos contained in the Political Speeches Dataset, can be interpreted as a step forward to the discovery of an approach generalisable over all kinds of videos. Therefore, it is advisable to continue the research and, especially, the data collection. In particular, for further research on highlights extraction from videos, it is advisable to choose video types that present a more explicit structure than the ones contained in the Political Speeches Dataset. This way, the investigation of appropriate Machine Learning architectures is not compromised by the intrinsic impossibility to understand the common patterns in these videos.

#### 8.3.2. ENLARGE THE POLITICAL SPEECHES DATASET

In order to improve the quality of the highlights extraction from the videos of the Political Speeches Dataset, it is a task for future work to expand the dataset with the inclusion of more speeches and more than one ground truth highlight clip per speech, so that the saliency labels would be more reliable. In addition, the availability of a larger amount of data is beneficial to Machine Learning algorithms.

#### 8.3.3. EXTEND THE MFN

Even though the MFN architecture, performing a complete multimodal features integration, resulted in lower accuracy than the multimodal approach based on features concatenation and Random Forests. This rejects the hypothesis that multimodal feature integration is superior to early multimodal features, that is feature concatenation. However, this result might be due to the fact that the standard MFN architecture is too shallow, and therefore, cannot produce high level representation of the features. In future work, the MFN model should be extended with a set of fully-connected or convolutional layers, in addition to the LSTM layers.

Another possibility is to try to use the MFN architecture on a new dataset for video summarisation or highlights extraction, that is provided with relevant labels.

#### 8.3.4. EXPAND THE CROWDSOURCING STUDY

Lastly, the crowdsourcing study presented in this thesis was limited by the time and monetary constraints. In future work, it would be interesting to conduct a more extensive crowdsourcing study, involving all the highlight clips produced in this project and a larger number of workers. This would allow to obtain a more thorough evaluation of the results of the proposed highlights extraction methods.

## BIBLIOGRAPHY

- [1] L.-P. Morency, R. Mihalcea, and P. Doshi, *Towards multimodal sentiment analysis: Harvesting opinions from the web*, in *Proceedings of the 13th international conference on multimodal interfaces* (ACM, 2011) pp. 169–176.
- [2] V. Pérez-Rosas, R. Mihalcea, and L.-P. Morency, *Utterance-level multimodal sentiment analysis*, in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1 (2013) pp. 973–982.
- [3] W. Zhang, Y. Zhang, L. Ma, J. Guan, and S. Gong, *Multimodal learning for facial expression recognition*, *Pattern Recognition* **48**, 3191 (2015).
- [4] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski, *et al.*, *Emonets: Multimodal deep learning approaches for emotion recognition in video*, *Journal on Multimodal User Interfaces* **10**, 99 (2016).
- [5] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, *Toward multimodal image-to-image translation*, in *Advances in Neural Information Processing Systems* (2017) pp. 465–476.
- [6] M. Merler, K.-N. C. Mac, D. Joshi, Q.-B. Nguyen, S. Hammer, J. Kent, J. Xiong, M. N. Do, J. R. Smith, and R. S. Feris, *Automatic curation of sports highlights using multimodal excitement features*, *IEEE Transactions on Multimedia* **21**, 1147 (2018).
- [7] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, *Human multimodal language in the wild: A novel dataset and interpretable dynamic fusion model*, in *Association for Computational Linguistics* (2018).
- [8] P. P. Liang, A. Zadeh, and L.-P. Morency, *Multimodal local-global ranking fusion for emotion recognition*, in *Proceedings of the 2018 on International Conference on Multimodal Interaction* (ACM, 2018) pp. 472–476.
- [9] X. Huang, M.-Y. Liu, S. Belongie, and J. Kautz, *Multimodal unsupervised image-to-image translation*, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018) pp. 172–189.
- [10] M. Burzo, M. Abouelenien, V. Perez-Rosas, and R. Mihalcea, *Multimodal deception detection*, .
- [11] J. R. Smith, D. Joshi, B. Huet, W. Hsu, and J. Cota, *Harnessing ai for augmenting creativity: Application to movie trailer creation*, in *Proceedings of the 2017 ACM on Multimedia Conference* (ACM, 2017) pp. 1799–1808.
- [12] A. Zadeh, P. P. Liang, N. Mazumder, S. Poria, E. Cambria, and L.-P. Morency, *Memory fusion network for multi-view sequential learning*, arXiv preprint arXiv:1802.00927 (2018).
- [13] M. T. Maybury, *Intelligent multimedia information retrieval* (Aaai Press, 1997).
- [14] M. R. Naphade, *On supervision and statistical learning for semantic multimedia analysis*, *Journal of Visual Communication and Image Representation* **15**, 348 (2004).
- [15] A. Mallik, P. Pasumarthi, and S. Chaudhury, *Multimedia ontology learning for automatic annotation and video browsing*, in *Proceedings of the 1st ACM international conference on Multimedia information retrieval* (ACM, 2008) pp. 387–394.
- [16] A. Divakaran, *Multimedia content analysis: theory and applications* (Springer Science & Business Media, 2009).
- [17] J. Zahálka and M. Worring, *Towards interactive, intelligent, and integrated multimedia analytics*, in *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)* (IEEE, 2014) pp. 3–12.

- [18] S. Dumais, E. Cutrell, J. J. Cadiz, G. Jancke, R. Sarin, and D. C. Robbins, *Stuff i've seen: a system for personal information retrieval and re-use*, in *Acm sigir forum*, Vol. 49 (ACM, 2016) pp. 28–35.
- [19] C. Manning, P. Raghavan, and H. Schütze, *Introduction to information retrieval*, *Natural Language Engineering* **16**, 100 (2010).
- [20] A. Berger and J. Lafferty, *Information retrieval as statistical translation*, in *ACM SIGIR Forum*, Vol. 51 (ACM, 2017) pp. 219–226.
- [21] R. Datta, D. Joshi, J. Li, and J. Z. Wang, *Image retrieval: Ideas, influences, and trends of the new age*, *ACM Computing Surveys (Csur)* **40**, 5 (2008).
- [22] Y. Liu, D. Zhang, G. Lu, and W.-Y. Ma, *A survey of content-based image retrieval with high-level semantics*, *Pattern recognition* **40**, 262 (2007).
- [23] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, *Neural codes for image retrieval*, in *European conference on computer vision* (Springer, 2014) pp. 584–599.
- [24] Z. Xiong, R. Radhakrishnan, and A. Divakaran, *Generation of sports highlights using motion activity in combination with a common audio feature extraction framework*, in *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, Vol. 1 (IEEE, 2003) pp. 1–5.
- [25] Z. Xiong, R. Radhakrishnan, A. Divakaran, and T. S. Huang, *Highlights extraction from sports video based on an audio-visual marker detection framework*, in *2005 IEEE International Conference on Multimedia and Expo* (IEEE, 2005) pp. 4–pp.
- [26] M. H. Kolekar and S. Sengupta, *Bayesian network-based customized highlight generation for broadcast soccer videos*, *IEEE Transactions on Broadcasting* **61**, 195 (2015).
- [27] T. Yao, T. Mei, and Y. Rui, *Highlight detection with pairwise deep ranking for first-person video summarization*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 982–990.
- [28] A. Tejero-de Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna, and E. Rahtu, *Summarization of user-generated sports video by using deep action recognition features*, *IEEE Transactions on Multimedia* **20**, 2000 (2018).
- [29] P. Mundur, Y. Rao, and Y. Yesha, *Keyframe-based video summarization using delaunay clustering*, *International Journal on Digital Libraries* **6**, 219 (2006).
- [30] P. Yousefi and L. I. Kuncheva, *Selective keyframe summarisation for egocentric videos based on semantic concept search*, in *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)* (IEEE, 2018) pp. 19–24.
- [31] K. Zhou and Y. Qiao, *Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward*, *arXiv preprint arXiv:1801.00054* (2017).
- [32] K. Zhou, T. Xiang, and A. Cavallaro, *Video summarisation by classification with deep reinforcement learning*, *arXiv preprint arXiv:1807.03089* (2018).
- [33] C. Gutwin, M. van der Kamp, M. S. Uddin, K. Stanley, I. Stavness, and S. Vail, *Improving early navigation in time-lapse video with spread loading*, (2019).
- [34] J. Carletta, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, W. Kraaij, M. Kronenthal, *et al.*, *The ami meeting corpus: A pre-announcement*, in *International workshop on machine learning for multimodal interaction* (Springer, 2005) pp. 28–39.
- [35] I. McCowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, *et al.*, *The ami meeting corpus*, in *Proceedings of the 5th International Conference on Methods and Techniques in Behavioral Research*, Vol. 88 (2005) p. 100.

- [36] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, *Introducing the recola multimodal corpus of remote collaborative and affective interactions*, in *2013 10th IEEE international conference and workshops on automatic face and gesture recognition (FG)* (IEEE, 2013) pp. 1–8.
- [37] A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. Morency, *Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph*, in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Vol. 1 (2018) pp. 2236–2246.
- [38] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, *Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos*, arXiv preprint arXiv:1606.06259 (2016).
- [39] F. Eyben, M. Wöllmer, and B. Schuller, *Openear—introducing the munich open-source emotion and affect recognition toolkit*, in *2009 3rd international conference on affective computing and intelligent interaction and workshops* (IEEE, 2009) pp. 1–6.
- [40] F. Eyben, M. Wöllmer, and B. Schuller, *Opensmile: the munich versatile and fast open-source audio feature extractor*, in *Proceedings of the 18th ACM international conference on Multimedia* (ACM, 2010) pp. 1459–1462.
- [41] D. Borth, T. Chen, R. Ji, and S.-F. Chang, *Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content*, in *Proceedings of the 21st ACM international conference on Multimedia* (ACM, 2013) pp. 459–460.
- [42] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, *Learning deep features for scene recognition using places database*, in *Advances in neural information processing systems* (2014) pp. 487–495.
- [43] *Google Scholar*, <https://scholar.google.it>, last accessed: 23 Sept 2019.
- [44] L. Yao, A. Torabi, K. Cho, N. Ballas, C. Pal, H. Larochelle, and A. Courville, *Describing videos by exploiting temporal structure*, in *Proceedings of the IEEE international conference on computer vision* (2015) pp. 4507–4515.
- [45] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, *Video summarization with long short-term memory*, in *European conference on computer vision* (Springer, 2016) pp. 766–782.
- [46] M. Rochan, L. Ye, and Y. Wang, *Video summarization using fully convolutional sequence networks*, arXiv preprint arXiv:1805.10538 (2018).
- [47] C. Gianluigi and S. Raimondo, *An innovative algorithm for key frame extraction in video summarization*, *Journal of Real-Time Image Processing* **1**, 69 (2006).
- [48] D. M. Russell, *A design pattern-based video summarization technique: moving from low-level signals to high-level structure*, in *System Sciences, 2000. Proceedings of the 33rd Annual Hawaii International Conference on* (IEEE, 2000) pp. 10–pp.
- [49] M. A. Smith and T. Kanade, *Video skimming for quick browsing based on audio and image characterization* (Citeseer, 1995).
- [50] S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, *Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method*, *Pattern Recognition Letters* **32**, 56 (2011).
- [51] M. Furini, F. Geraci, M. Montangero, and M. Pellegrini, *Stimo: Still and moving video storyboard for the web scenario*, *Multimedia Tools and Applications* **46**, 47 (2010).
- [52] D. Potapov, M. Douze, Z. Harchaoui, and C. Schmid, *Category-specific video summarization*, in *European conference on computer vision* (Springer, 2014) pp. 540–555.
- [53] B. Gong, W.-L. Chao, K. Grauman, and F. Sha, *Diverse sequential subset selection for supervised video summarization*, in *Advances in Neural Information Processing Systems* (2014) pp. 2069–2077.

- [54] J. Luo, C. Papin, and K. Costello, *Towards extracting semantically meaningful key frames from personal video clips: from humans to computers*, IEEE Transactions on Circuits and Systems for Video Technology **19**, 289 (2009).
- [55] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*, in *Advances in neural information processing systems* (2012) pp. 1097–1105.
- [56] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, *Learning spatiotemporal features with 3d convolutional networks*, in *Proceedings of the IEEE international conference on computer vision* (2015) pp. 4489–4497.
- [57] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, *Creating summaries from user videos*, in *European conference on computer vision* (Springer, 2014) pp. 505–520.
- [58] K. Zhou, Y. Qiao, and T. Xiang, *Deep reinforcement learning for unsupervised video summarization with diversity-representativeness reward*, in *Thirty-Second AAAI Conference on Artificial Intelligence* (2018).
- [59] P. J. Rousseeuw and L. Kaufman, *Finding groups in data*, Series in Probability & Mathematical Statistics 199034 (1), 111 (1990).
- [60] Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, *Tvsum: Summarizing web videos using titles*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 5179–5187.
- [61] W.-S. Chu, Y. Song, and A. Jaimes, *Video co-summarization: Video summarization by visual co-occurrence*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015) pp. 3584–3592.
- [62] Y. Zhang, M. Kampffmeyer, X. Liang, D. Zhang, M. Tan, and E. P. Xing, *Dtr-gan: Dilated temporal relational adversarial network for video summarization*, arXiv preprint arXiv:1804.11228 (2018).
- [63] J. Long, E. Shelhamer, and T. Darrell, *Fully convolutional networks for semantic segmentation*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 3431–3440.
- [64] B. Shahraray, *Scene change detection and content-based sampling of video sequences*, in *Digital Video Compression: Algorithms and Technologies 1995*, Vol. 2419 (International Society for Optics and Photonics, 1995) pp. 2–13.
- [65] L. Baraldi, C. Grana, and R. Cucchiara, *Shot and scene detection via hierarchical clustering for re-using broadcast video*, in *International Conference on Computer Analysis of Images and Patterns* (Springer, 2015) pp. 801–811.
- [66] M. Gygli, *Ridiculously fast shot boundary detection with fully convolutional neural networks*, in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)* (IEEE, 2018) pp. 1–4.
- [67] H. Jiang, T. Lin, and H. Zhang, *Video segmentation with the support of audio segmentation and classification*, in *Proc. IEEE ICME* (2000).
- [68] J. P. Campbell, *Speaker recognition: A tutorial*, Proceedings of the IEEE **85**, 1437 (1997).
- [69] D. Castán, D. Tavarez, P. Lopez-Otero, J. Franco-Pedroso, H. Delgado, E. Navas, L. Docio-Fernández, D. Ramos, J. Serrano, A. Ortega, et al., *Albayzín-2014 evaluation: audio segmentation and classification in broadcast news domains*, EURASIP Journal on Audio, Speech, and Music Processing **2015**, 33 (2015).
- [70] T. Butko, C. Nadeu Camprubí, and H. Schulz, *Albayzin-2010 audio segmentation evaluation: evaluation setup and results*, in *VI Jornadas en Tecnología del Habla and II Iberian SLTech Workshop* (2010) pp. 305–308.
- [71] F. Font, G. Roma, and X. Serra, *Freesound technical demo*, in *Proceedings of the 21st ACM international conference on Multimedia* (ACM, 2013) pp. 411–412.
- [72] G. Hu, *100 non-speech environmental sounds*, <http://www.cse.ohio-state.edu/dwang/pnl/corpus/HuCorpus.html>, last accessed: 23 Sept 2019.



- [73] A. Bietti, F. Bach, and A. Cont, *An online em algorithm in hidden (semi-) markov models for audio segmentation and clustering*, in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on* (IEEE, 2015) pp. 1881–1885.
- [74] O. Cappé, *Online em algorithm for hidden markov models*, *Journal of Computational and Graphical Statistics* **20**, 728 (2011).
- [75] R. M. Neal and G. E. Hinton, *A view of the em algorithm that justifies incremental, sparse, and other variants*, in *Learning in graphical models* (Springer, 1998) pp. 355–368.
- [76] D. Giannoulis, E. Benetos, D. Stowell, M. Rossignol, M. Lagrange, and M. D. Plumbley, *Detection and classification of acoustic scenes and events: An ieee aasp challenge*, in *2013 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics* (IEEE, 2013) pp. 1–4.
- [77] N. Keriven, D. Garreau, and I. Poli, *Newma: a new method for scalable model-free online change-point detection*, arXiv preprint arXiv:1805.08061 (2018).
- [78] H. McGurk and J. MacDonald, *Hearing lips and seeing voices*, *Nature* **264**, 746 (1976).
- [79] Y. Mroueh, E. Marcheret, and V. Goel, *Deep multimodal learning for audio-visual speech recognition*, in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, 2015) pp. 2130–2134.
- [80] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, *Encara2: Real-time detection of multiple faces at different resolutions in video streams*, *Journal of visual communication and image representation* **18**, 130 (2007).
- [81] M. Tapaswi, Y. Zhu, R. Stiefelhagen, A. Torralba, R. Urtasun, and S. Fidler, *Movieqa: Understanding stories in movies through question-answering*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2016) pp. 4631–4640.
- [82] X. Chen and C. L. Zitnick, *Learning a recurrent visual representation for image caption generation*, arXiv preprint arXiv:1411.5654 (2014).
- [83] L. Yu, E. Park, A. C. Berg, and T. L. Berg, *Visual madlibs: Fill in the blank description generation and question answering*, in *Proceedings of the IEEE international conference on computer vision* (2015) pp. 2461–2469.
- [84] A. Agrawal, J. Lu, S. Antol, M. Mitchell, C. L. Zitnick, D. Parikh, and D. Batra, *Vqa: Visual question answering*, *International Journal of Computer Vision* **123**, 4 (2017).
- [85] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, arXiv preprint arXiv:1301.3781 (2013).
- [86] R. Kiros, Y. Zhu, R. R. Salakhutdinov, R. Zemel, R. Urtasun, A. Torralba, and S. Fidler, *Skip-thought vectors*, in *Advances in neural information processing systems* (2015) pp. 3294–3302.
- [87] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, *Going deeper with convolutions*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015) pp. 1–9.
- [88] K. Chen, T. Bui, C. Fang, Z. Wang, and R. Nevatia, *Amc: Attention guided multi-modal correlation learning for image search*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017) pp. 2644–2652.
- [89] B. Korbar, D. Tran, and L. Torresani, *Scsampler: Sampling salient clips from video for efficient action recognition*, arXiv preprint arXiv:1904.04289 (2019).
- [90] A. Vinciarelli, M. Pantic, and H. Bourlard, *Social signal processing: Survey of an emerging domain*, *Image and vision computing* **27**, 1743 (2009).
- [91] C. Shan, S. Gong, and P. W. McOwan, *Facial expression recognition based on local binary patterns: A comprehensive study*, *Image and vision Computing* **27**, 803 (2009).

- [92] X. Zhu and D. Ramanan, *Face detection, pose estimation, and landmark localization in the wild*, in *2012 IEEE conference on computer vision and pattern recognition* (IEEE, 2012) pp. 2879–2886.
- [93] J. Yang, K. Wang, X. Peng, and Y. Qiao, *Deep recurrent multi-instance learning with spatio-temporal features for engagement intensity prediction*, in *Proceedings of the 2018 on International Conference on Multimodal Interaction* (ACM, 2018) pp. 594–598.
- [94] B. Amos, B. Ludwiczuk, M. Satyanarayanan, *et al.*, *Openface: A general-purpose face recognition library with mobile applications*, *CMU School of Computer Science* **6** (2016).
- [95] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, *Hand keypoint detection in single images using multiview bootstrapping*, in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (2017) pp. 1145–1153.
- [96] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Zadeh, and L.-P. Morency, *Efficient low-rank multimodal fusion with modality-specific factors*, arXiv preprint arXiv:1806.00064 (2018).
- [97] Y. Song, L.-P. Morency, and R. Davis, *Action recognition by hierarchical sequence summarization*, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013) pp. 3562–3569.
- [98] S. S. Rajagopalan, L.-P. Morency, T. Baltrušaitis, and R. Goecke, *Extending long short-term memory for multi-view structured learning*, in *European Conference on Computer Vision* (Springer, 2016) pp. 338–353.
- [99] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, *Multimodal machine learning: A survey and taxonomy*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2018).
- [100] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, *Microsoft coco: Common objects in context*, in *European conference on computer vision* (Springer, 2014) pp. 740–755.
- [101] M. J. Huiskes, B. Thomee, and M. S. Lew, *New trends and ideas in visual concept detection: the mir flickr retrieval evaluation initiative*, in *Proceedings of the international conference on Multimedia information retrieval* (ACM, 2010) pp. 527–536.
- [102] X. Chen, H. Fang, T.-Y. Lin, R. Vedantam, S. Gupta, P. Dollár, and C. L. Zitnick, *Microsoft coco captions: Data collection and evaluation server*, arXiv preprint arXiv:1504.00325 (2015).
- [103] J. Hessel, D. Mimno, and L. Lee, *Quantifying the visual concreteness of words and topics in multimodal datasets*, arXiv preprint arXiv:1804.06786 (2018).
- [104] P. Pasupat and P. Liang, *Compositional semantic parsing on semi-structured tables*, arXiv preprint arXiv:1508.00305 (2015).
- [105] G. McKeown, M. F. Valstar, R. Cowie, and M. Pantic, *The semaine corpus of emotionally coloured character interactions*, in *2010 IEEE International Conference on Multimedia and Expo* (IEEE, 2010) pp. 1079–1084.
- [106] J. Pennington, R. Socher, and C. Manning, *Glove: Global vectors for word representation*, in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (2014) pp. 1532–1543.
- [107] iMotions, *Emotient Facial Expression Analysis Engine*, <https://imotions.com/emotient/>, last accessed: 23 Sept 2019.
- [108] G. Degottex, J. Kane, T. Drugman, T. Raitio, and S. Scherer, *Covarep—a collaborative voice analysis repository for speech technologies*, in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on* (IEEE, 2014) pp. 960–964.
- [109] J. Howe, *The rise of crowdsourcing*, *Wired magazine* **14**, 1 (2006).
- [110] E. Estellés-Arolas and F. González-Ladrón-De-Guevara, *Towards an integrated crowdsourcing definition*, *Journal of Information science* **38**, 189 (2012).

- [111] J. Howe, *Crowdsourcing: How the power of the crowd is driving the future of business* (Random House, 2008).
- [112] F. Oliveira, I. Ramos, and L. Santos, *Definition of a crowdsourcing innovation service for the european smes*, in *International Conference on Web Engineering* (Springer, 2010) pp. 412–416.
- [113] M. Vukovic, M. Lopez, and J. Laredo, *Peoplecloud for the globally integrated enterprise*, in *Service-Oriented Computing. ICSOC/ServiceWave 2009 Workshops* (Springer, 2010) pp. 109–114.
- [114] M. N. Wexler, *Reconfiguring the sociology of the crowd: exploring crowdsourcing*, *International Journal of Sociology and Social Policy* **31**, 6 (2011).
- [115] M. Buhrmester, T. Kwang, and S. D. Gosling, *Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?* *Perspectives on psychological science* **6**, 3 (2011).
- [116] G. Paolacci, J. Chandler, and P. G. Ipeirotis, *Running experiments on amazon mechanical turk*, *Judgment and Decision making* **5**, 411 (2010).
- [117] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, *Imagenet: A large-scale hierarchical image database*, in *2009 IEEE conference on computer vision and pattern recognition* (Ieee, 2009) pp. 248–255.
- [118] G. A. Sigurdsson, G. Varol, X. Wang, A. Farhadi, I. Laptev, and A. Gupta, *Hollywood in homes: Crowdsourcing data collection for activity understanding*, in *European Conference on Computer Vision* (Springer, 2016) pp. 510–526.
- [119] D. McDuff, R. El Kaliouby, and R. Picard, *Crowdsourced data collection of facial responses*, in *Proceedings of the 13th international conference on multimodal interfaces* (ACM, 2011) pp. 11–18.
- [120] M. Soleymani and M. Larson, *Crowdsourcing for affective annotation of video: Development of a viewer-reported boredom corpus*, (2010).
- [121] R. M. McCreddie, C. Macdonald, and I. Ounis, *Crowdsourcing a news query classification dataset*, in *Proceedings of the ACM SIGIR 2010 workshop on crowdsourcing for search evaluation (CSE 2010)* (ACM Geneva, Switzerland, 2010) pp. 31–38.
- [122] M. Post, C. Callison-Burch, and M. Osborne, *Constructing parallel corpora for six indian languages via crowdsourcing*, in *Proceedings of the Seventh Workshop on Statistical Machine Translation* (Association for Computational Linguistics, 2012) pp. 401–409.
- [123] O. F. Zaidan and C. Callison-Burch, *Crowdsourcing translation: Professional quality from non-professionals*, in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (Association for Computational Linguistics, 2011) pp. 1220–1229.
- [124] C. Sun, N. Rampalli, F. Yang, and A. Doan, *Chimera: Large-scale classification using machine learning, rules, and crowdsourcing*, *Proceedings of the VLDB Endowment* **7**, 1529 (2014).
- [125] A. Khosla, R. Hamid, C.-J. Lin, and N. Sundaresan, *Large-scale video summarization using web-image priors*, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013) pp. 2698–2705.
- [126] O. Alonso and R. Baeza-Yates, *Design and implementation of relevance assessments using crowdsourcing*, in *Advances in Information Retrieval*, edited by P. Clough, C. Foley, C. Gurrin, G. J. F. Jones, W. Kraaij, H. Lee, and V. Mudoch (Springer Berlin Heidelberg, Berlin, Heidelberg, 2011) pp. 153–164.
- [127] T. Caselli and O. Inel, *Crowdsourcing storylines: Harnessing the crowd for causal relation annotation*, in *Proceedings of the Workshop Events and Stories in the News 2018* (2018) pp. 44–54.
- [128] A. G. S. de Herrera, A. Foncubierta-Rodríguez, D. Markonis, R. Schaer, and H. Müller, *Crowdsourcing for medical image classification*, *Swiss Medical Informatics* **30** (2014).
- [129] C. Heipke, *Crowdsourcing geospatial data*, *ISPRS Journal of Photogrammetry and Remote Sensing* **65**, 550 (2010).

- [130] A. E. Ainasoja, A. Hietanen, J. Lankinen, and J.-K. Kämäräinen, *Keyframe-based video summarization with human in the loop*, in *VISIGRAPP* (2018).
- [131] R. Mihalcea and P. Tarau, *Textrank: Bringing order into text*, in *Proceedings of the 2004 conference on empirical methods in natural language processing* (2004).
- [132] F. Barrios, F. López, L. Argerich, and R. Wachenchauser, *Variations of the similarity function of textrank for automated summarization*, arXiv preprint arXiv:1602.03606 (2016).
- [133] M. Buckland and F. Gey, *The relationship between recall and precision*, *Journal of the American society for information science* **45**, 12 (1994).
- [134] *Figure Eight*, <https://www.figure-eight.com>, last accessed: 23 Sept 2019.
- [135] C.-Y. Lin, *Rouge: A package for automatic evaluation of summaries*, Text Summarization Branches Out (2004).
- [136] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency, *Openface 2.0: Facial behavior analysis toolkit*, in *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)* (IEEE, 2018) pp. 59–66.
- [137] S. Wold, K. Esbensen, and P. Geladi, *Principal component analysis*, *Chemometrics and intelligent laboratory systems* **2**, 37 (1987).
- [138] P. Jha, *A Brief Overview of Loss Functions in Pytorch*, <https://medium.com/udacity-pytorch-challenges/a-brief-overview-of-loss-functions-in-pytorch-c0ddb78068f7>, last accessed: 23 Sept 2019.
- [139] M. A. Tanner and W. H. Wong, *The calculation of posterior distributions by data augmentation*, *Journal of the American statistical Association* **82**, 528 (1987).
- [140] E. Duffin, *State minimum wage rates in the united states as of january 1, 2019, by state ( in u.s. dollars)*, <https://www.statista.com/statistics/238997/minimum-wage-by-us-state/>.
- [141] A. Carvalho, S. Dimitrov, and K. Larson, *How many crowdsourced workers should a requester hire?* *Annals of Mathematics and Artificial Intelligence* **78**, 45 (2016).
- [142] J. Bai and P. Perron, *Computation and analysis of multiple structural change models*, *Journal of applied econometrics* **18**, 1 (2003).
- [143] H. N. Boone and D. A. Boone, *Analyzing likert data*, *Journal of extension* **50**, 1 (2012).



# POLITICAL SPEECHES DATASET

## A.1. FULL LIST OF POLITICAL SPEECHES FROM THE DATASET

The following list contains the speeches that form the Political Speeches Dataset. Each speech is identified by a unique title. There are listed three URLs for each speeches: the first is the link to the full speech video, the second is the link to the highlight clip, preceded by the source, and the third link refers to the official transcript.

- Barack Obama

1. Barack Obama's farewell speech

- <https://www.youtube.com/watch?v=udrKnXueTWO>
- CNN <https://www.youtube.com/watch?v=Zd4sS1WInuo&t=169s>
- Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamafarewelladdress.htm>

2. 2004 Democratic National Convention Keynote Address

- <https://www.youtube.com/watch?v=ueMNqdB1QIE>
- CNN <https://www.youtube.com/watch?v=yJzjyYL815Y>
- Transcript <https://www.americanrhetoric.com/speeches/convention2004/barackobama2004dnc.htm>

3. Former President Obama unleashes on Trump

- <https://www.youtube.com/watch?v=sHAKDTlv8fA>
- NBC <https://www.youtube.com/watch?v=cPRPpyc20Bw>
- Transcript <https://www.vox.com/policy-and-politics/2018/9/7/17832024/obama-speech-trump-illinois-transcript>

4. Barack Obama's Presidential Announcement

- [https://www.youtube.com/watch?time\\_continue=5&v=gdJ7Ad15WCA](https://www.youtube.com/watch?time_continue=5&v=gdJ7Ad15WCA)
- NBC <https://www.youtube.com/watch?v=Ybaw8A3jti0>
- Transcript <https://www.americanrhetoric.com/speeches/barackobamacandidacyforresident.htm>

5. South Carolina Democratic Primary Victory Speech

- [https://www.youtube.com/watch?time\\_continue=3&v=4BobS7RjS2E](https://www.youtube.com/watch?time_continue=3&v=4BobS7RjS2E)
- FOX <https://www.youtube.com/watch?v=NDG-9sEMGbE>
- Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamasouthcarolinavictory.htm>

6. A More Perfect Union

- <https://www.youtube.com/watch?v=zrp-v2tHaDo>

- CNN <https://www.youtube.com/watch?v=y1BpPWpoRVE>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobamaperfectunion.htm>
7. Democratic Nomination Victory Speech
- [https://www.youtube.com/watch?time\\_continue=65&v=dtL-1V30Z0c](https://www.youtube.com/watch?time_continue=65&v=dtL-1V30Z0c)
  - CBS [https://www.youtube.com/watch?v=\\_0ep0JNpKag&t=22s](https://www.youtube.com/watch?v=_0ep0JNpKag&t=22s)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobamademocraticnominationvictoryspeech.htm>
8. President-Elect Victory Speech
- [https://www.youtube.com/watch?time\\_continue=61&v=HfHbw3n0EIM](https://www.youtube.com/watch?time_continue=61&v=HfHbw3n0EIM)
  - USATODAY [https://www.youtube.com/watch?v=19Du4vI\\_noQ](https://www.youtube.com/watch?v=19Du4vI_noQ)
  - Transcript <https://www.americanrhetoric.com/speeches/convention2008/barackobamavictoryspeech.htm>
9. Address in Strasbourg Town Hall
- [https://www.youtube.com/watch?time\\_continue=1519&v=6Ez01U-VLnA](https://www.youtube.com/watch?time_continue=1519&v=6Ez01U-VLnA)
  - TPM [https://www.youtube.com/watch?v=T17qt\\_tU6wo&t=123s](https://www.youtube.com/watch?v=T17qt_tU6wo&t=123s)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamasstrasbourgspeech.htm>
10. Speech at Hradcany Square in Prague
- [https://www.youtube.com/watch?v=\\_lcpg6yQ0Yw](https://www.youtube.com/watch?v=_lcpg6yQ0Yw)
  - EURACTIV <https://www.youtube.com/watch?v=6gW8x6Tp8sU>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamapraguespeech.htm>
11. Speech to the Turkish Parliament
- <https://www.youtube.com/watch?v=x3PrM9WJZus>
  - CNN <https://www.youtube.com/watch?v=gNJJ3FS7f-g>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaturkishparliament.htm>
12. Speech to the American Medical Association
- [https://www.youtube.com/watch?time\\_continue=36&v=TTFzVY9qyQc](https://www.youtube.com/watch?time_continue=36&v=TTFzVY9qyQc)
  - CBS <https://www.youtube.com/watch?v=40DYxBZYhlg>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaamericanmedicalassociation.htm>
13. Address to the Ghanaian Parliament
- [https://www.youtube.com/watch?time\\_continue=73&v=CYvwYWabWvs](https://www.youtube.com/watch?time_continue=73&v=CYvwYWabWvs)
  - VOANews <https://www.youtube.com/watch?v=9miWVxe2Wa8>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaghanaparliament.htm>
14. Speech to a Joint Session of Congress on Health Care Reform
- [https://www.youtube.com/watch?time\\_continue=27&v=SSJugLU5M58](https://www.youtube.com/watch?time_continue=27&v=SSJugLU5M58)
  - APArchive <https://www.youtube.com/watch?v=QU61vik-s00>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamajointsessionhealthcare.htm>
15. Shanghai Town Hall With Future Chinese Leaders
- [https://www.youtube.com/watch?v=NrQ-Vj0n\\_KQ](https://www.youtube.com/watch?v=NrQ-Vj0n_KQ)
  - VOANews <https://www.youtube.com/watch?v=UtY2ZiP9N60>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamashanghaihall.htm>

16. Remarks and Press Conference on the Gulf Oil Spill Disaster
  - [https://www.youtube.com/watch?time\\_continue=25&v=2rT7IANtSjo](https://www.youtube.com/watch?time_continue=25&v=2rT7IANtSjo)
  - AP <https://www.youtube.com/watch?v=ZN0i04R8mCY>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamagulfoilspillpresser.htm>
17. Speech on the National Wireless Initiative
  - [https://www.youtube.com/watch?time\\_continue=627&v=9ZkuafwQplo](https://www.youtube.com/watch?time_continue=627&v=9ZkuafwQplo)
  - AP <https://www.youtube.com/watch?v=AyswL5PS3xM>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamanationalwirelessinitiative.htm>
18. American Energy Security Address at Georgetown University
  - [https://www.youtube.com/watch?time\\_continue=16&v=HpRTtfmXXLY](https://www.youtube.com/watch?time_continue=16&v=HpRTtfmXXLY)
  - VOA News <https://www.youtube.com/watch?v=wjvUBDWjWrU>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaamericanenergysecurity.htm>
19. Announcement of the Death of Osama Bin Laden
  - <https://www.youtube.com/watch?v=ZNYmK19-d0U>
  - EuroNews <https://www.youtube.com/watch?v=ER7Txs0atWM>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamasabinladendeath.htm>
20. Commencement Speech at Booker T. Washington HS
  - [https://www.youtube.com/watch?time\\_continue=35&v=fZCmwxpwnI](https://www.youtube.com/watch?time_continue=35&v=fZCmwxpwnI)
  - CBS <https://www.youtube.com/watch?v=nf6f85MJ6Bk>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamabookertwashington.htm>
21. Address to the British Parliament
  - [https://www.youtube.com/watch?time\\_continue=26&v=fp85zRg2cwg](https://www.youtube.com/watch?time_continue=26&v=fp85zRg2cwg)
  - AP Archive <https://www.youtube.com/watch?v=IMR0SB--CPg>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamabritishparliament.htm>
22. Speech at Memorial for Joplin Tornado Victims
  - [https://www.youtube.com/watch?time\\_continue=955&v=qrmjhK\\_PUmQ](https://www.youtube.com/watch?time_continue=955&v=qrmjhK_PUmQ)
  - AP <https://www.youtube.com/watch?v=j6siPa2hibg>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamajoplinmemorial.htm>
23. Afghanistan Troop Reduction Address to the Nation
  - [https://www.youtube.com/watch?time\\_continue=25&v=ai01D82uBs8](https://www.youtube.com/watch?time_continue=25&v=ai01D82uBs8)
  - The Telegraph <https://www.youtube.com/watch?v=E5oFORSliuM>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaafghanistanwardrawdown.htm>
24. Address at the Martin Luther King Memorial Dedication
  - [https://www.youtube.com/watch?time\\_continue=62&v=QR8GEDjT-x4](https://www.youtube.com/watch?time_continue=62&v=QR8GEDjT-x4)
  - CBS <https://www.youtube.com/watch?v=XihGOVT66VI>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamamlkmemorialdedication.htm>
25. Speech on Ending the War in Iraq Responsibly
  - [https://www.youtube.com/watch?time\\_continue=2&v=G9Z7tdukQuo](https://www.youtube.com/watch?time_continue=2&v=G9Z7tdukQuo)

- TheTelegraph <https://www.youtube.com/watch?v=ykeQajHR0w4>
  - Transcript <https://www.americanrhetoric.com/speeches/wariniraq/barackobamaendingiraqwar.htm>
26. Address at an Associated Press Luncheon
- [https://www.youtube.com/watch?time\\_continue=53&v=49-tKE-Ka2Y](https://www.youtube.com/watch?time_continue=53&v=49-tKE-Ka2Y)
  - APArchive [https://www.youtube.com/watch?v=wC18c\\_FLjps](https://www.youtube.com/watch?v=wC18c_FLjps)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaapluncheon.htm>
27. On the USSC Ruling on the Affordable Care Act
- [https://www.youtube.com/watch?time\\_continue=40&v=b5zU1y\\_0Geo](https://www.youtube.com/watch?time_continue=40&v=b5zU1y_0Geo)
  - CNN <https://www.youtube.com/watch?v=re3U-Rbct90>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamausshealthcareruling.htm>
28. Second Democratic Presidential Nomination Acceptance Speech
- [https://www.youtube.com/watch?time\\_continue=130&v=cXp3ksU3QoE](https://www.youtube.com/watch?time_continue=130&v=cXp3ksU3QoE)
  - ABCNews <https://www.youtube.com/watch?v=Zw-ec4grvvc>
  - Transcript <https://www.americanrhetoric.com/speeches/convention2012/barackobama2012dnc.htm>
29. Address at the Transfer of Remains Ceremony for the Victims of the Attacks on the U.S. Consulate at Benghazi
- [https://www.youtube.com/watch?time\\_continue=23&v=bkVhQRpMSOQ](https://www.youtube.com/watch?time_continue=23&v=bkVhQRpMSOQ)
  - TheTelegraph <https://www.youtube.com/watch?v=asxqp9NF3v0>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamabenghazivictims.htm>
30. Address at the Clinton Global Initiative on Human Trafficking
- [https://www.youtube.com/watch?time\\_continue=65&v=2rz5\\_eg-dZY](https://www.youtube.com/watch?time_continue=65&v=2rz5_eg-dZY)
  - WSJ <https://www.youtube.com/watch?v=e4DnpRI-5ds>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaclintonglobalinitiative2012.htm>
31. 67th Session of the United Nations General Assembly Address
- [https://www.youtube.com/watch?time\\_continue=5&v=-GqYCKV2wzA](https://www.youtube.com/watch?time_continue=5&v=-GqYCKV2wzA)
  - CBSNews <https://www.youtube.com/watch?v=K9zXZc608eY>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaunitednations67.htm>
32. Second Presidential Election Victory Speech
- [https://www.youtube.com/watch?time\\_continue=143&v=nv9NwKAjmt0](https://www.youtube.com/watch?time_continue=143&v=nv9NwKAjmt0)
  - ABCNews <https://www.youtube.com/watch?v=-r7WZgc-2Qw>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamasecondpresidentialvictoryspeech.htm>
33. Address at Yangon University
- [https://www.youtube.com/watch?time\\_continue=1&v=PWNK6x0c000](https://www.youtube.com/watch?time_continue=1&v=PWNK6x0c000)
  - WSJ <https://www.youtube.com/watch?v=HtA1MhcdLsE>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamayangonuniversity.htm>
34. Statement on the Sandy Hook Elementary School Shootings in Newtown, Connecticut
- [https://www.youtube.com/watch?v=\\_q-\\_T87j1MY](https://www.youtube.com/watch?v=_q-_T87j1MY)
  - ABCNews <https://www.youtube.com/watch?v=a0LgGYr2M6g>



- Transcript <https://www.npr.org/2012/12/16/167412995/transcript-president-Obama-at-sandy-hook-prayer-vigil?t=1557731185579>
35. Interfaith Prayer Vigil Address at Newtown High School
- [https://www.youtube.com/watch?time\\_continue=49&v=YmalVRadC78](https://www.youtube.com/watch?time_continue=49&v=YmalVRadC78)
  - AP [https://www.youtube.com/watch?v=i3LT\\_6MpudI](https://www.youtube.com/watch?v=i3LT_6MpudI)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamanewtownvigilspeech.htm>
36. Second Presidential Inaugural Address
- [https://www.youtube.com/watch?time\\_continue=10&v=agKFUaf74bA](https://www.youtube.com/watch?time_continue=10&v=agKFUaf74bA)
  - CNN <https://www.youtube.com/watch?v=JuZZpQ5LLOw&t=73s>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamascondinauguraladdress.htm>
37. Address on Comprehensive Immigration Reform
- [https://www.youtube.com/watch?time\\_continue=2&v=51VIuW8vJ\\_E](https://www.youtube.com/watch?time_continue=2&v=51VIuW8vJ_E)
  - CBSNews [https://www.youtube.com/watch?v=\\_1Y1IpbCYmc&t=2s](https://www.youtube.com/watch?v=_1Y1IpbCYmc&t=2s)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobaimmigrationreform.htm>
38. Speech on Sequester Impact at Newport News Shipbuilding
- [https://www.youtube.com/watch?time\\_continue=7&v=GefYjSb6RZc](https://www.youtube.com/watch?time_continue=7&v=GefYjSb6RZc)
  - WSJ [https://www.youtube.com/watch?v=EBD-CUph\\_f4](https://www.youtube.com/watch?v=EBD-CUph_f4)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamasquesternewportnews.htm>
39. Morehouse College Commencement Address
- [https://www.youtube.com/watch?time\\_continue=15&v=e50Tt9qJRQk](https://www.youtube.com/watch?time_continue=15&v=e50Tt9qJRQk)
  - CBSNews <https://www.youtube.com/watch?v=6Up-RpTo4mY>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamamorehousecollegecommencement.htm>
40. Address on Drones and Terrorism at the National Defense University
- [https://www.youtube.com/watch?time\\_continue=34&v=fEnUbwXAof0](https://www.youtube.com/watch?time_continue=34&v=fEnUbwXAof0)
  - CNN <https://www.youtube.com/watch?v=JAUTikhuyFc>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamanationaldefenseuniversity.htm>
41. Address to the People of Northern Ireland
- [https://www.youtube.com/watch?time\\_continue=523&v=sc9gupTbsIo](https://www.youtube.com/watch?time_continue=523&v=sc9gupTbsIo)
  - APArchive [https://www.youtube.com/watch?v=YNTbw\\_fkWvA](https://www.youtube.com/watch?v=YNTbw_fkWvA)
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamanorthernireland.htm>
42. Address on the Economy at Knox College
- [https://www.youtube.com/watch?time\\_continue=6&v=6CzBA1UPdC8](https://www.youtube.com/watch?time_continue=6&v=6CzBA1UPdC8)
  - CNN <https://www.youtube.com/watch?v=IKK-0M-YaKI>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaknoxcollegeeconomy.htm>
43. Address at the 'Let Freedom Ring' Ceremony Commemorating the 50th Anniversary of the March on Washington, D.C.
- [https://www.youtube.com/watch?time\\_continue=53&v=M-wEk1lmZMo](https://www.youtube.com/watch?time_continue=53&v=M-wEk1lmZMo)
  - Euronews <https://www.youtube.com/watch?v=jk7gKFANMvC>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamalletfreedomringceremony.htm>

44. Statement on the U.S. Government Shutdown
  - [https://www.youtube.com/watch?time\\_continue=26&v=HmRA\\_tML2tE](https://www.youtube.com/watch?time_continue=26&v=HmRA_tML2tE)
  - TheGuardian <https://www.youtube.com/watch?v=2390ScRIQKE>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamagovernmentsshutdown.htm>
45. Veterans Day Address
  - [https://www.youtube.com/watch?time\\_continue=45&v=psIn1A-jkGM](https://www.youtube.com/watch?time_continue=45&v=psIn1A-jkGM)
  - TheTelegraph <https://www.youtube.com/watch?v=FplmoLHH0cU>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaveteransday2013.htm>
46. Memorial Address for Nelson Mandela
  - [https://www.youtube.com/watch?time\\_continue=6&v=Sgg0sfjsL0c](https://www.youtube.com/watch?time_continue=6&v=Sgg0sfjsL0c)
  - WSJ <https://www.youtube.com/watch?v=SgtXy1vTcGY>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamanelsonmandelamemorial.htm>
47. On the Outcome of U.S. Intelligence Programs Review
  - [https://www.youtube.com/watch?v=Z5Zk2Fy\\_KDI](https://www.youtube.com/watch?v=Z5Zk2Fy_KDI)
  - TheTelegraph <https://www.youtube.com/watch?v=2QESZ9HoKIE>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamasignalsintelreview.htm>
48. United States Military Academy Commencement Address
  - [https://www.youtube.com/watch?time\\_continue=5&v=fG\\_hX\\_XM4Ks](https://www.youtube.com/watch?time_continue=5&v=fG_hX_XM4Ks)
  - Bloomberg <https://www.youtube.com/watch?v=0N00-kP2sse>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamawestpointcommencement2014.htm>
49. 70th Anniversary of D-Day Address
  - [https://www.youtube.com/watch?time\\_continue=15&v=ewcQ9hCP9rM](https://www.youtube.com/watch?time_continue=15&v=ewcQ9hCP9rM)
  - CNN <https://www.youtube.com/watch?v=04HCqqsrfJc>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaday70.htm>
50. Statement on the Downing Malaysia Airlines Flight 17
  - [https://www.youtube.com/watch?time\\_continue=11&v=RNVOBkDsYzM](https://www.youtube.com/watch?time_continue=11&v=RNVOBkDsYzM)
  - CNN <https://www.youtube.com/watch?v=I3ucKjzPmna>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaflightmh17.htm>
51. On Authorizing Targeted Air Strikes and Humanitarian Aid in Iraq
  - [https://www.youtube.com/watch?list=UUyxR1FDqcWM4y7FfpiAN3KQ&time\\_continue=247&v=ax4a6cH1Wjs](https://www.youtube.com/watch?list=UUyxR1FDqcWM4y7FfpiAN3KQ&time_continue=247&v=ax4a6cH1Wjs)
  - ABCNews <https://www.youtube.com/watch?v=tqi-04ajNVY>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamairaqairstrikehumanitarianeffort.htm>
52. Address to the Nation on the Islamic State of Iraq and the Levant
  - [https://www.youtube.com/watch?time\\_continue=10&v=KvRd17vXaXM](https://www.youtube.com/watch?time_continue=10&v=KvRd17vXaXM)
  - CNN <https://www.youtube.com/watch?v=pwp8qKvE-0g>
  - Transcript <https://www.americanrhetoric.com/speeches/barackobama/barackobamaispeechtonation.htm>

- Donald Trump

53. Donald Trump Announces Presidential Campaign
  - <https://www.youtube.com/watch?v=SpMJx0-Hy0M>
  - WSJ <https://www.youtube.com/watch?v=bx6V-e2DQW0>
  - Transcript <http://time.com/3923128/donald-trump-announcement-speech/>
54. Donald Trump Holds a Political Rally in Green Bay, Wisconsin
  - <https://www.youtube.com/watch?v=GY7W4jQJgc4>
  - CBSNews [https://www.youtube.com/watch?v=X90\\_tKLKBe8](https://www.youtube.com/watch?v=X90_tKLKBe8)
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-green-bay-wisconsin-april-27-2019>
55. Donald Trump Addresses a National Republican Campaign Committee Fundraiser
  - <https://www.youtube.com/watch?v=TT16ZXAA2cc>
  - AP <https://www.youtube.com/watch?v=twG43zzeipY>
  - Transcript <https://factba.se/transcript/donald-trump-speech-rncc-fundraiser-dinner-april-2-2019>
56. Donald Trump Holds a Political Rally in Grand Rapids, Michigan - March 28, 2019
  - <https://www.youtube.com/watch?v=GorMdb8k-Mg>
  - GlobalNews <https://www.youtube.com/watch?v=tXS13pcyMCs>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-grand-rapids-march-28-2019>
57. Donald Trump Holds a Political Rally in El Paso, Texas - February 11, 2019
  - [https://www.youtube.com/watch?time\\_continue=7&v=02HrvfGn8JE](https://www.youtube.com/watch?time_continue=7&v=02HrvfGn8JE)
  - AP <https://www.youtube.com/watch?v=mtffJvi0Fo8>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-el-paso-february-11-2019>
58. Donald Trump Delivers the State of the Union
  - <https://www.youtube.com/watch?v=JTyp6hbCka0>
  - AP <https://www.youtube.com/watch?v=b0luKryMXc0>
  - Transcript <https://factba.se/transcript/donald-trump-speech-state-of-the-union-february-5-2019>
59. Donald Trump Addresses the Nation from the Oval Office on the Border Wall
  - [https://www.youtube.com/watch?v=LF\\_xB\\_gg63U](https://www.youtube.com/watch?v=LF_xB_gg63U)
  - OneAmericaNewsNetwork <https://www.youtube.com/watch?v=bDguqD8iD58>
  - Transcript <https://factba.se/transcript/donald-trump-speech-oval-office-immigration-january-8-2019>
60. Donald Trump Speaks to Troops at Al Asad Air Base in Iraq
  - <https://www.youtube.com/watch?v=s1Ae0v9VTpk>
  - GuardianNews <https://www.youtube.com/watch?v=sXw1v8cnPVY>
  - Transcript <https://factba.se/transcript/donald-trump-speech-troops-holiday-visit-iraq-december-26-2018>
61. Donald Trump Holds a Political Rally in Pensacola, Florida
  - <https://www.youtube.com/watch?v=1yM-ylsYD1g>
  - FoxNews <https://www.youtube.com/watch?v=bkk6fGtt3vA>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-pensacola-fl-november-3-2018>
62. Donald Trump Holds a Political Rally in Huntington, West Virginia
  - <https://www.youtube.com/watch?v=iXkMkzVnfiE>
  - CBSNews <https://www.youtube.com/watch?v=Cb1cZs6fCj0>

- Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-huntington-wv-november-2-2018>
63. Donald Trump Holds a Political Rally in Murphysboro, Illinois
- <https://www.youtube.com/watch?v=7riALaGzlw>
  - ABC7Chicago <https://www.youtube.com/watch?v=0ThDha9lBb8>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-murphysboro-il-october-27-2018>
64. Donald Trump Holds a Political Rally in Houston, Texas
- <https://www.youtube.com/watch?v=7ceTnNsMw-M>
  - NBCNews <https://www.youtube.com/watch?v=fkevCJ2NJok>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-houston-tx-october-22-2018>
65. Donald Trump Holds a Political Rally in Richmond, Kentucky
- [https://www.youtube.com/watch?v=6WTBFEz\\_BUg](https://www.youtube.com/watch?v=6WTBFEz_BUg)
  - FoxNews <https://www.youtube.com/watch?v=qDn45M7uo1A>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-richmond-ky-october-13-2018>
66. Donald Trump Holds a Political Rally in Wheeling, West Virginia
- <https://www.youtube.com/watch?v=Zjpu7q-YDGI>
  - FoxNews <https://www.youtube.com/watch?v=c5FkHuwROEc&t=6s>
  - Transcript <https://factba.se/transcript/donald-trump-speech-maga-rally-wheeling-wv-september-29-2018>
67. Donald Trump Addresses the 73rd Session of the United Nations
- [https://www.youtube.com/watch?v=FRYgy\\_IL-8s](https://www.youtube.com/watch?v=FRYgy_IL-8s)
  - WashingtonPost <https://www.youtube.com/watch?v=oUnnNZ2n1Nk>
  - Transcript <https://factba.se/transcript/donald-trump-speech-un-general-assembly-september-25-2018>
68. Donald Trump Holds a Political Rally in Las Vegas, Nevada
- <https://www.youtube.com/watch?v=s8MDTtudmZk>
  - OneAmericaNewsNetwork <https://www.youtube.com/watch?v=fnbib1Itml8>
  - Transcript <https://factba.se/transcript/donald-trump-speech-political-rally-maga-las-vegas-september-20-2018>
69. Donald Trump Speaks at a GOP Fundraiser in Sioux Falls, South Dakota
- [https://www.youtube.com/watch?v=2BA46U\\_rcDE](https://www.youtube.com/watch?v=2BA46U_rcDE)
  - APArchive <https://www.youtube.com/watch?v=s5wUHQ59KE8>
  - Transcript <https://factba.se/transcript/donald-trump-speech-gop-fundraiser-sioux-falls-south-dakota-september-7-2018>
70. Donald Trump Signs a Retirement Savings Executive Order in Charlotte, NC
- [https://www.youtube.com/watch?time\\_continue=1&v=tvKzV\\_U1Q\\_0](https://www.youtube.com/watch?time_continue=1&v=tvKzV_U1Q_0)
  - APArchive <https://www.youtube.com/watch?v=okallcUdjEs>
  - Transcript <https://factba.se/transcript/donald-trump-speech-retirement-savings-charlotte-nc-august-31-2018>
71. Donald Trump Addresses Ohio Republican Party State Dinner
- <https://www.youtube.com/watch?v=6X9VTsiG124>
  - WCMH-TV <https://www.youtube.com/watch?v=bHKZESYBDvI>
  - Transcript <https://factba.se/transcript/donald-trump-speech-ohio-gop-state-dinner-august-24-2018>

72. Donald Trump Holds a Political Rally in West Columbia, South Carolina
  - <https://www.youtube.com/watch?v=B-ve11P63Vg>
  - AP <https://www.youtube.com/watch?v=iegYGcs6n0k>
  - Transcript <https://factba.se/transcript/donald-trump-speech-south-carolina-gop-mcmaster-june-23-2018>
73. Donald Trump Holds a Political Rally in Moon Township, Pennsylvania
  - <https://www.youtube.com/watch?v=dUte8CdssxU>
  - FOXNews <https://www.youtube.com/watch?v=UqDZyURB9yM>
  - Transcript <https://factba.se/transcript/donald-trump-speech-rally-saccone-pennsylvania-march-10-2018>
74. Donald Trump Delivers the State of the Union Address
  - <https://www.youtube.com/watch?v=8JqPD2njvNI>
  - FOXNews <https://www.youtube.com/watch?v=4G0einedJTQ>
  - Transcript <https://factba.se/transcript/donald-trump-speech-state-of-the-union-january-30-2018>
75. Donald Trump Addresses the World Economic Forum
  - <https://www.youtube.com/watch?v=da0CyzcJ2o8>
  - WashingtonPost <https://www.youtube.com/watch?v=HCs30AS4Xy4>
  - Transcript <https://factba.se/transcript/donald-trump-speech-world-economic-forum-davos-january-26-2018>
76. Donald Trump Holds a Political Rally in Pensacola, Florida December 8, 2017
  - <https://www.youtube.com/watch?v=7m4D5Lc-xgM>
  - FOXNews <https://www.youtube.com/watch?v=6ewzQp9BKMY&t=202s>
  - Transcript <https://factba.se/transcript/donald-trump-speech-make-america-great-again-pensacola-december-8-2017>
77. Donald Trump Addresses Iran Strategy
  - <https://www.youtube.com/watch?v=cKqP0q8Ah9g>
  - AP [https://www.youtube.com/watch?v=NEao\\_NxHpjA](https://www.youtube.com/watch?v=NEao_NxHpjA)
  - Transcript <https://factba.se/transcript/donald-trump-speech-iran-strategy-october-13-2017>
78. Donald Trump Delivers a Speech on Tax Reform in Harrisburg, PA
  - <https://www.youtube.com/watch?v=vZpEtOfNuKE>
  - WashingtonPost <https://www.youtube.com/watch?v=VFX77vuxRlU>
  - Transcript <https://factba.se/transcript/donald-trump-speech-tax-reform-harrisburg-october-11-2017>
79. Donald Trump Addresses the United Nations General Assembly
  - <https://www.youtube.com/watch?v=PRp2j8iN99E>
  - WashingtonPost <https://www.youtube.com/watch?v=dc3H3cYy0ns>
  - Transcript <https://factba.se/transcript/donald-trump-speech-united-nations-general-assembly-september-19-2017>
80. Donald Trump Holds a Political Rally in Huntington, West Virginia August 3, 2017
  - <https://www.youtube.com/watch?v=y9CaEMgLC8s>
  - WashingtonPost <https://www.youtube.com/watch?v=Uo0DivHn4>
  - Transcript <https://factba.se/transcript/donald-trump-speech-rally-huntington-wv-august-3-2017>
81. Donald Trump Holds a Political Rally in Youngstown, Ohio
  - <https://www.youtube.com/watch?v=AgNa3zRFmEM>

- WashingtonPost <https://www.youtube.com/watch?v=r5xsPxGHolc&t=12s>
  - Transcript <https://factba.se/transcript/donald-trump-speech-rally-youngstown-ohio-july-25-2017>
82. Donald Trump Commissions the USS Gerald R. Ford in Norfolk, Virginia
- <https://www.youtube.com/watch?v=Gu7QYafUNEg>
  - AP <https://www.youtube.com/watch?v=PVwgCfZvElc>
  - Transcript <https://factba.se/transcript/donald-trump-speech-uss-gerald-r-for-d-commissioning-july-22-2017>
83. Donald Trump Delivers a Speech in Krasiski Square in Warsaw, Poland
- [https://www.youtube.com/watch?time\\_continue=3&v=zfa4ZdhN6aQ](https://www.youtube.com/watch?time_continue=3&v=zfa4ZdhN6aQ)
  - NBCNews <https://www.youtube.com/watch?v=eEfKKw1-h8c>
  - Transcript <https://factba.se/transcript/donald-trump-speech-warsaw-poland-july-6-2017>
84. Donald Trump at Celebrate Freedom Rally
- <https://www.youtube.com/watch?v=Do4g7D44Uqg>
  - WashingtonPost <https://www.youtube.com/watch?v=FWTQg14VSEY>
  - Transcript <https://factba.se/transcript/donald-trump-speech-celebrate-freedom-rally-july-1-2017>
85. Donald Trump Speaks at the Department of Transportation
- <https://www.youtube.com/watch?v=Ni0JIbIT05c>
  - APArchive <https://www.youtube.com/watch?v=B78zsydAnU8>
  - Transcript <https://factba.se/transcript/donald-trump-speech-transportation-regulation-june-9-2017>
86. Donald Trump Announces Withdrawal From Paris Climate Accord
- <https://www.youtube.com/watch?v=2Ew7a0kx9gM>
  - BBC <https://www.youtube.com/watch?v=jP55meWlLt4>
  - Transcript <https://factba.se/transcript/donald-trump-speech-paris-climate-accord-june-1-2017>
87. Donald Trump at the Memorial Day Ceremony at Arlington Cemetery
- [https://www.youtube.com/watch?time\\_continue=10&v=x\\_sLRaS1KhU](https://www.youtube.com/watch?time_continue=10&v=x_sLRaS1KhU)
  - AP <https://www.youtube.com/watch?v=V-hqHKZYCUI>
  - Transcript <https://factba.se/transcript/donald-trump-speech-memorial-day-arlington-may-29-2017>
88. Donald Trump Receives an Honorary Doctorate at Liberty University
- <https://www.youtube.com/watch?v=2XgwTFH3agk>
  - APArchive <https://www.youtube.com/watch?v=4EErvmJlD-4>
  - Transcript <https://factba.se/transcript/donald-trump-commencement-address-liberty-university-may-13-2017>
89. Donald Trump Addresses a Joint Session of Congress
- [https://www.youtube.com/watch?time\\_continue=1&v=pp3AsAxdKbQ](https://www.youtube.com/watch?time_continue=1&v=pp3AsAxdKbQ)
  - VOA [https://www.youtube.com/watch?v=T\\_Q8aIeKzeU](https://www.youtube.com/watch?v=T_Q8aIeKzeU)
  - Transcript <https://factba.se/transcript/donald-trump-speech-congress-february-28-2017>
90. Donald Trump - Department of Homeland Security, January 25, 2017
- <https://www.youtube.com/watch?v=5ZMHgvT9liU>
  - AP <https://www.youtube.com/watch?v=eLcucpAwrfM>
  - Transcript <https://factba.se/transcript/donald-trump-speech-washington-dc-january-25-2017>

- George W. Bush
  91. Eulogy for Father, George H.W. Bush
    - <https://www.youtube.com/watch?v=RrRzED2vOHA>
    - NBCNews [https://www.youtube.com/watch?v=z4\\_ntSr301g](https://www.youtube.com/watch?v=z4_ntSr301g)
    - Transcript <https://www.americanrhetoric.com/speeches/gwbuseulogyforfather.htm>
  92. Address at the Dedication of the National Museum of African-American History & Culture
    - <https://www.youtube.com/watch?v=sLRuyBd9fRQ>
    - DailyMail <https://www.youtube.com/watch?v=nAIUa7vuWE>
    - Transcript <https://www.americanrhetoric.com/speeches/gwbushafricanamericanmuseum.htm>
  93. Farewell Address to the Nation
    - [https://www.youtube.com/watch?time\\_continue=235&v=g-NK1EKmcX8](https://www.youtube.com/watch?time_continue=235&v=g-NK1EKmcX8)
    - VOANews <https://www.youtube.com/watch?v=JsB4iUPzEyA>
    - Transcript <https://www.americanrhetoric.com/speeches/gwbushfarewelladdress.htm>
  94. Announces End of Major Combat Operations in Iraq
    - [https://www.youtube.com/watch?time\\_continue=436&v=5yCsmwoMecU](https://www.youtube.com/watch?time_continue=436&v=5yCsmwoMecU)
    - APArchive <https://www.youtube.com/watch?v=v-MHEMPqvaI>
    - Transcript <https://www.americanrhetoric.com/speeches/wariniraq/gwbushiraq5103.htm>
- Hillary Clinton
  95. Presidential Campaign Concession Speech
    - <https://www.youtube.com/watch?v=WPWPW5FtXCc>
    - TheGlobeAndMail <https://www.youtube.com/watch?v=9JeQpJHD88E>
    - Transcript <https://www.americanrhetoric.com/speeches/hillaryclinton2016campaignconcession.htm>
  96. Democratic Presidential Nomination Acceptance
    - <https://www.youtube.com/watch?v=C6GnHBEBWYE>
    - TheGlobeAndMail <https://www.youtube.com/watch?v=D13vtwAQNos>
    - Transcript <https://www.americanrhetoric.com/speeches/convention2016/hillaryclinton2016dnc.htm>
  97. International Human Rights Day Address at Palais des Nations
    - [https://www.youtube.com/watch?time\\_continue=4&v=WIqynW5EbIQ](https://www.youtube.com/watch?time_continue=4&v=WIqynW5EbIQ)
    - CNN <https://www.youtube.com/watch?v=iwf8PQjPz9o>
    - Transcript <https://www.americanrhetoric.com/speeches/hillaryclintonintlbgthumanrights.htm>
- James B. Comey
  98. Georgetown University Speech on Race and Law Enforcement
    - [https://www.youtube.com/watch?time\\_continue=1354&v=sbx4HAm6Rc8](https://www.youtube.com/watch?time_continue=1354&v=sbx4HAm6Rc8)
    - CBSN <https://www.youtube.com/watch?v=qFdrHo8kQ9k>
    - Transcript <https://www.americanrhetoric.com/speeches/jamescomeygeorgetownraceandlaw.htm>
- Nikki Haley
  99. Speech Announcing U.S. Withdrawal from the United Nations Human Rights Council
    - <https://www.youtube.com/watch?v=xls30enLyeg>
    - DailyMail <https://www.youtube.com/watch?v=enPATSVFFa0>
    - Transcript <https://www.americanrhetoric.com/speeches/nikkyhaleyunhumanrightscouncilwithdrawal.htm>



(a) Speech 1



(b) Speech 31



(c) Speech 66



(d) Speech 74



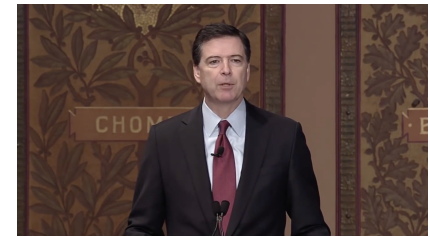
(e) Speech 94



(f) Speech 95



(g) Speech 96



(h) Speech 98

Figure A.1: Example of videos from the dataset

## A.2. MISSING WORDS USING "COMMON CRAWL" FOR THE PRE-TRAINED WORD VECTORS

The pre-trained GloVe vector "Common Crawl" (840B tokens)<sup>1</sup> was used to derive the text features for the Political Speeches dataset. Here is the list of all the 428 words that could not be represented, as they were not present in the pre-trained vector.

- |                  |                      |                  |                  |
|------------------|----------------------|------------------|------------------|
| 1. 10point       | 10. acostas          | 19. alexeis      | 28. antilynching |
| 2. 18003083515   | 11. activeduty       | 20. allfemale    | 29. antimedia    |
| 3. 18003182596   | 12. africanamericans | 21. allpowerful  | 30. antipolice   |
| 4. 22hour        | 13. agendasetting    | 22. alqaedas     | 31. antiu        |
| 5. 45year        | 14. aidsfree         | 23. alshabaab    | 32. antiworker   |
| 6. 49year        | 15. alassad          | 24. alwaki       | 33. aplusplus    |
| 7. 5story        | 16. alassads         | 25. americanmade | 34. aqap         |
| 8. 70year        | 17. alawites         | 26. anticoal     | 35. aqaps        |
| 9. abdulmutallab | 18. albaghdadi       | 27. antilaw      | 36. arakis       |

<sup>1</sup>Available at <https://nlp.stanford.edu/projects/glove/>



37. areareare	83. crimeinfested	129. feinsteins	175. honesttogod
38. arminarm	84. crossstrait	130. fiveandahalf	176. horsford
39. asianamerican	85. crossstraits	131. fivetime	177. hultgren
40. asiapacific	86. culberson	132. foreignborn	178. illconceived
41. assads	87. daffaires	133. fortynine	179. illintentioned
42. ataturks	88. dahlberg	134. fourandahalf	180. inslee
43. aumf	89. daugaard	135. fourmonthold	181. interamerican
44. azars	90. davino	136. fouryearold	182. intermediaterange
45. barnea	91. debras	137. frenchled	183. isil
46. battlehardened	92. decadeslong	138. freshfaced	184. israelipalestinian
47. battlescarred	93. deeprooted	139. fudan	185. jakari
48. battletested	94. deficitneutral	140. fuefficient	186. jcpoa
49. bedfordstuyvesant	95. deficitreduction	141. gaetz	187. jeffress
50. bedstuy	96. degioia	142. georgetowns	188. jimmers
51. bergdahl	97. democracys	143. getzs	189. jindals
52. bestequipped	98. democratcontrolled	144. ghanas	190. jobcrushing
53. bestled	99. demoss	145. gillum	191. jobkilling
54. besttrained	100. denisha	146. gillums	192. jongun
55. betos	101. dermatend	147. giveandtake	193. joplins
56. billiondollar	102. dewine	148. godgiven	194. judahs
57. blumenauer	103. dingell	149. gohmert	195. kavanaughs
58. bottomup	104. dminus	150. goodfaith	196. kentuckys
59. bprd	105. doortodoor	151. goodpaying	197. keplinger
60. brabzzz	106. dotorg	152. gordonreed	198. kiner
61. bruster	107. duerson	153. gorsuch	199. kooser
62. bunche	108. duncanson	154. gossage	200. largertanlife
63. byrum	109. eema	155. governmentapproved	201. lateterm
64. cabella	110. eightyyearold	156. governmentrun	202. latorre
65. cancerfree	111. eisenhowers	157. greatgranddaughter	203. lauraandsteve
66. carnegies	112. employerbased	158. greatlooking	204. liberalleaning
67. casebycase	113. epas	159. griefstricken	205. lisam
68. cashstrapped	114. etchasketch	160. grothman	206. lochner
69. catapulter	115. evatt	161. halfcentury	207. longestrunning
70. catchandrelease	116. everwidening	162. handinhand	208. lordstown
71. cavador	117. excellencys	163. hardesthit	209. lubic
72. cavanaugh	118. exportoriented	164. hardhit	210. madiba
73. cedillo	119. expresidents	165. hawthornes	211. madibas
74. centurieslong	120. eyetoeye	166. healthcaregov	212. majoritys
75. charleshenri	121. ezeiels	167. henryoswald	213. makeorbreak
76. cleanburning	122. fl8s	168. hesburgh	214. malias
77. cleareyed	123. falwells	169. highestever	215. mandelas
78. commanderinchief	124. familyowned	170. highestrated	216. marigo
79. congresss	125. farleft	171. highpaying	217. marketbased
80. cosel	126. fastmoving	172. highvalue	218. masaryk
81. covino	127. fbis	173. hivaid	219. masserias
82. craziestlooking	128. feeforservice	174. hochsprung	220. maximumsecurity
			221. mcchystals
			222. medicareforall
			223. medicaregov
			224. meritbased

225. mh17	277. pattons	329. selfimposed	381. thirdgeneration
226. mickens	278. pelosis	330. selfinflicted	382. thirdrate
227. middleincome	279. pepfar	331. selfreflection	383. thirtyseven
228. milania	280. pettus	332. selfreflective	384. threeandahalf
229. militaryindustrial	281. pointscoring	333. selfreliance	385. threefifths
230. mingalaba	282. pointtopoint	334. selfreliant	386. threemonth
231. minuchin	283. polands	335. seongho	387. threepoint
232. missourians	284. postworld	336. seventyeight	388. threeyearold
233. mitchs	285. prenew	337. sevenyearold	389. tibbetts
234. mogai	286. presidentelect	338. severson	390. timetested
235. moneygall	287. presidentforlife	339. shalonda	391. tongji
236. msms	288. prestons	340. shantz	392. toughestever
237. multimilliondollar	289. proamerican	341. sherlach	393. toughoncrime
238. mulvaney	290. proconstitution	342. shinseki	394. trilliondollar
239. muslimmajority	291. profamily	343. shouldertoshoulder	395. trinace
240. mymymy	292. progrowth	344. shulkin	396. trinitys
241. naftas	293. projobs	345. shuttlesworth	397. tudela
242. nagornokarabakh	294. prorussian	346. shwedagon	398. twofifths
243. nakooma	295. proworker	347. sicom	399. ukraines
244. nantz	296. ptaff	348. simpsonbowles	400. waistdeep
245. nativeborn	297. publicprivate	349. singlepayer	401. waitll
246. natoma	298. qaedas	350. singleyear	402. walzwerk
247. natos	299. quoteunquote	351. sinjar	403. warmbier
248. naypyidaw	300. raciallycharged	352. slashandburn	404. warroad
249. nicelooking	301. raishin	353. slavka	405. weaponsgrade
250. nineoneone	302. randhir	354. slowwalking	406. webberley
251. ninetyfive	303. raqqa	355. smeeting	407. wellconnected
252. ninetyseven	304. recordsetting	356. smokefilled	408. wellcrafted
253. ninetythree	305. remsburg	357. snowdens	409. wellfunctioning
254. nineyearold	306. renacci	358. sociallyconscious	410. wellmanaged
255. nogami	307. renees	359. soontobe	411. wellmeaning
256. noncall	308. renforth	360. sotloff	412. welloff
257. nonpolitician	309. rigell	361. splitsecond	413. wellpaying
258. nosetonose	310. risktakers	362. spurofthemoment	414. wellversed
259. nrcc	311. rockribbed	363. stabenow	415. wellworn
260. nucleararmed	312. rohingya	364. stanshall	416. wenjian
261. nypds	313. runofthemill	365. stateled	417. winnefeld
262. obamacares	314. russianbacked	366. statesponsored	418. winnertakeall
263. odierno	315. sarek	367. steinle	419. woolston
264. officerinvolved	316. scalise	368. stoltzenberg	420. wroe
265. oilrich	317. schmucker	369. sungo	421. yazidi
266. onefifth	318. schumers	370. surfacetoair	422. yazidis
267. oneinamillion	319. schwerner	371. susies	423. yemenis
268. onesixth	320. searchandrescue	372. sutley	424. yemens
269. onestory	321. secondclass	373. swearengin	425. ygi
270. oroarke	322. secondguessed	374. sworn	426. zerosum
271. outbuild	323. secondquarter	375. syrias	427. zerotolerance
272. outeducate	324. secretarygeneral	376. taneys	428. zinke
273. outinnovate	325. selfdefeating	377. tarkanian	
274. outofcontrol	326. selfdescribed	378. taxandspend	
275. outofpocket	327. selfexecuting	379. tenminute	
276. parasteel	328. selfgovernment	380. tenyearold	

# B

## CROWDSOURCING STUDY

### B.1. HIGHLIGHTS PRESENTED IN THE EVALUATIONS

From the highlight clips created for this thesis, some were selected and included in the two crowdsourcing jobs employed to obtain the human evaluation. The chosen highlight clips were extracted from five of speeches contained in the Political Speeches Dataset. These highlights are all visible on YouTube and can be found from the urls available in the following list.

#### 23 Barack Obama - "Afghanistan Troop Reduction Address to the Nation"

Ground truth [https://www.youtube.com/embed/rm\\_y1F4RBb8](https://www.youtube.com/embed/rm_y1F4RBb8)

Baseline [https://www.youtube.com/embed/Ub\\_f0uRrOLM](https://www.youtube.com/embed/Ub_f0uRrOLM)

Random Forest <https://www.youtube.com/embed/oFGhv75qhh0>

MFN <https://www.youtube.com/embed/88cSAi2COWY>

#### 26 Barack Obama - "Address at an Associated Press Luncheon"

Ground truth <https://www.youtube.com/embed/LSFurVmHyFw>

Baseline <https://www.youtube.com/embed/WgM0b-I3ML0>

Random Forest <https://www.youtube.com/embed/OB6RZyKxMoU>

MFN [https://www.youtube.com/embed/hvV\\_zk5n8YI](https://www.youtube.com/embed/hvV_zk5n8YI)

#### 46 Barack Obama - "Memorial Address for Nelson Mandela"

Ground truth [https://www.youtube.com/embed/VPt0xi0h0\\_Q](https://www.youtube.com/embed/VPt0xi0h0_Q)

Baseline <https://www.youtube.com/embed/gSyMEIh50fw>

Random Forest <https://www.youtube.com/embed/pf2kSIsq8zQ>

MFN <https://www.youtube.com/embed/5TNjqgeBG1Q>

#### 61 Donald Trump - "Donald Trump Holds a Political Rally in Pensacola Florida"

Ground truth <https://www.youtube.com/embed/xd0bVUNDR1o>

Baseline <https://www.youtube.com/embed/Rnc5cKnihP0>

Random Forest <https://www.youtube.com/embed/dzLfrvqzfm0>

MFN <https://www.youtube.com/embed/VYB90Nm8SXs>

#### 82 Donald Trump - "Donald Trump Commissions the USS Gerald R Ford in Norfolk Virginia"

Ground truth [https://www.youtube.com/embed/ENPT\\_dkbS3I](https://www.youtube.com/embed/ENPT_dkbS3I)

Baseline <https://www.youtube.com/embed/BC6bAf04jME>


Random Forest <https://www.youtube.com/embed/7QCz-5mTcrE>

MFN <https://www.youtube.com/embed/1s5NL930qYs>

## B.2. FIGURE EIGHT SURVEYS

This section reports an example of what the evaluation tasks published on the crowdsourcing platform Figure Eight [134] looked like. Figure B.1 corresponds to one task from the individual assessment survey, while figure B.2 to one task from the pairwise comparison survey.

Highlight clip



73

Watch later Share

1) *This highlight clip gives me information about what the full speech was about.*

How much do you agree with this statement? (required)

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Why did you make this choice? (required)

2) *This highlight clip makes me want to watch the full speech (independently on my political preference or my interest in politics).*

How much do you agree with this statement? (required)

Strongly disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree
<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Why did you make this choice? (required)

3) *I like the politician in the video.*

Answer: (required)

Yes

No

4) *I have already heard this speech.*

Answer: (required)

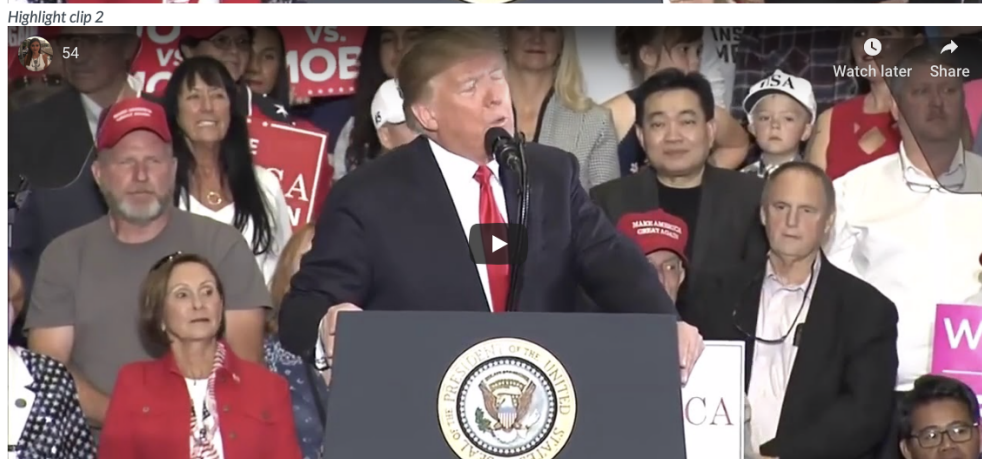
Yes

No

---

Did you find it difficult to answer these questions? Why or why not? (required)

Figure B.1: Example of one task from the individual assessment survey



1a) Which video gives more information about the topic of the full speech? (required)

- Highlight clip 1
- Highlight clip 2

1b) Why did you make this choice? (required)

2a) Which video contains more repetitions of the same sentences? (required)

- Highlight clip 1
- Highlight clip 2
- The two clips contain the same amount of repetitions

2b) Why did you make this choice? (required)

3a) Which video is more engaging (independently on its content or the political tendency)? (required)

- Highlight clip 1
- Highlight clip 2

3b) Why did you make this choice? (required)

4a) Which video is more cinematic? Namely, which highlight clip possesses the qualities and characteristics of a television production? (required)

- Highlight clip 1
- Highlight clip 2

4b) Why did you make this choice? (required)

5a) In which video is the speaker more expressive? (required)

- Highlight clip 1
- Highlight clip 2

5b) Why did you make this choice? (required)

Did you find it difficult to answer these questions? Why or why not? (required)

Figure B.2: Example of one task from the pairwise comparison survey