1290

UNIVERSIDADE D
COIMBRA

João Pedro Mendes Gonçalves

SENTIMENT ANALYSIS IN SOCIAL NETWORKS

VOLUME 1

Dissertação no âmbito do Mestrado de Sistemas Inteligentes orientada pelas Professoras Doutoras Catarina Helena Branco Simões da Silva, Bernardete Martins Ribeiro e Joana Madeira Martins Costa e apresentada ao Departamento de Engenharia Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra

Setembro de 2021

João Pedro Mendes Gonçalves

Sentiment analysis in social networks

Dissertação no âmbito do Mestrado em Engenharia Informática
orientada pelas Professoras Doutoras Catarina Silva, Bernardete Ribeiro e
Joana Madeira Martins Costa e apresentada ao Departamento de Engenharia
Informática da Faculdade de Ciências e Tecnologia da Universidade de Coimbra.

Setembro de 2021

Acknowledgments

Obrigado aos os meus pais, que sempre me apoiaram financeiramente e pessoalmente e sem os quais o meu percurso académico não era possível. Obrigado às minhas orientadoras, pelo apoio, pelas críticas constructivas que sempre me deram e por fomentarem o meu espiríto crítico, que foi uma mais valia ao longo do desenvolvimento deste trabalho. Finalmente, um obrigado aos amigos e ao resto da família que sempre me apoiaram ao longo deste semestre.

Abstract

Nowadays, social media plays an important role in connecting people all around the world. The information shared on these platforms is freely available and may be used to assess public opinion and to resolve challenges faced by groups and individuals. However, there are many uses of social media. Due to this, platforms with a focus in promoting discussion about present events, public figures and brands are preferred in the literature (e.g Twitter), as this data is useful for sentiment analysis.

The content shared on social media often contains multiple types of data, such as text, images and videos. Certain platforms (e.g Twitter) are more focused on textual content while others (e.g Instagram, pinterest) are mainly based on visual content. This presents a new challenge for sentiment analysis, as in this case it is necessary to classify multiple forms of data at once.

In this work, the main objective is to develop models with good capabilities for performing multimodal sentiment analysis on social media data. For this purpose, deep learning based approaches are tested and implemented, using models such as Long Short Term Memory Neural Networks and Convolutional Neural Networks. Other methods to increase performance are also considered, such as feature extraction from objects in images.

Given the initial lack of good quality datasets, data was collected to build a new dataset. However, during this work, a new high quality dataset was made available and is used instead.

Table of contents

1 Introduction

1.1 Motivation and challenges

The invention of the world wide web, 30 years ago, allowed people from all over the globe to interact with each other from anywhere, at any given time. As the internet spread, the world became increasingly connected. More recently, the rise of social media managed to further accelerate this phenomenon by simplifying the way we communicate over the world wide web.

Social media are the technologies which allow users to share information, opinions, ideas, and more through computers, smartphones and other electronic devices. It is currently one of the most popular ways to spend time online and its growth is not stopping anytime soon. According to the company Alexa, who analyzes Internet traffic, 2 of the 3 most visited non-chinese websites are social media, with Youtube and Facebook being the 2nd and 7th most visited, respectively (Alexa, 2020). When looking at their monthly users it is easy to understand why. Facebook has the most monthly users with 2.701 billion, followed by Youtube with 2 billion. Other well known platforms include Instagram with 1.158 billion users and Twitter with 353 million (Datareportal, 2021).

Even though in the early days of the Internet online interaction was mostly about sending text messages nowadays there are a lot more possibilities including sharing images, videos, organizing events, playing game, etc. Considering the fact that a lot of this information is publicly shared it is possible to get easy access to massive amounts of data in the form of text, images, videos and others.

The content shared online can be quite useful because it allows the understanding of the opinions of the public on various subjects in real time and many different entities can benefit from this knowledge. For example, brands with a large presence in the market may be mentioned daily by hundreds or even thousands of costumers. Using human resources to monitor large amounts of activity is difficult and automatic systems can help this process. Many researchers are studying sentiment analysis, which consists of the extraction of sentiment from various forms of data (e.g text, images, video) and identifying the emotions or polarity, e.g., positive, negative, neutral, present. Information from multiple modalities can also be used together, referred to as multimodal sentiment analysis.

However, there are major challenges in this field. While textual sentiment analysis has been widely studied, other modalities, i.e., images and video, have not been sufficiently researched and, therefore, model performance is subpar. Secondly, in the case of multimodal analysis, it is unclear what the best strategy is for fusing multiple modalities. Furthermore, the lack of adequately labeled multimodal datasets is also an issue since it hinders the performance and reliability of models.

1.2 Objectives

The primary goal of this work is to implement various models with the capabilities to correctly classify the polarity of a textual and visual social media dataset. To solve the challenges presented in the previous sections, the proposed objectives are:

- Study techniques for text and image analysis
- Improve sentiment classification of textual and visual data
- Test the developed methods on a high quality dataset

1.3 Outline

This report is divided into 7 chapters.

The first one describes how social media affects us, the motivation and challenges of sentiment analysis, and the objectives of this work.

Chapter 2 describes feature extraction, followed by the main approaches used and a brief discussion. Afterwards, some available datasets are detailed and a conclusion is presented.

The next chapter describes the proposed approach, starting by identifying the dataset used and how data is preprocessed. Subsequently, strategies for handling multiple annotators and choosing models and their parameters are detailed. Then, data sampling is explained and metrics are presented.

After this, the next chapter presents the results by comparing them to other works, followed by a brief discussion.

Finally, in the last chapter, the main takeaways from the work done are explained, an overview is given, and future work is discussed.

2 State of the art

In this chapter, the state of the art will be discussed. First, techniques for extracting meaningful features will be presented. Afterwards, different approaches for classifying data is discussed. Finally, at the end of each section, a brief discussion will be presented comparing different methods. Each section is dedicated to a different type of data which is commonly analyzed (text, images, multimodal and others) and will follow the same format.

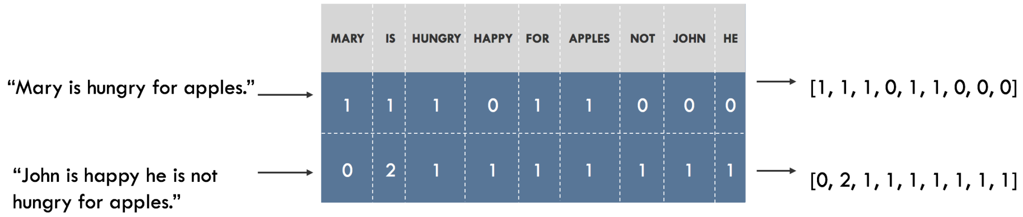
2.1 Text analysis

2.1.1 Feature extraction

There are many factors that can affect the sentiment present in text, including the words used, context, punctuation and others. Therefore, it is important to extract meaningful features since many approaches involve the use of models that require them. However, since there is often noise in the data (e.g., URLs, HTML tags, ...) preprocessing is a very important first step. This is especially necessary in social media based datasets which are written in an informal manner and thus are extra prone to noise with factors like misspellings, abbreviations and others being common.

One of the key factors to determine sentiment are the words used. Some, such as love and awful can directly affect how a sentence is perceived while others like not, very and little can act as modifiers, changing and/or negating the polarity of other words, i.e., how positive/negative they are. Lexicon-based algorithms make use of dictionaries that assign a polarity score to each word. In the case of machine learning this feature extraction step becomes more difficult because the models used need numeric and fixed-size data. To solve this problem there are two main types of techniques used: the bag-of-words model and word embeddings.

Bag-of-words (Figure 1): This model works by mapping each unique word to an integer that can then be used as an index in a vector. This vector can represent features such as word counts or frequencies. An example of a frequency method is term frequency–inverse document frequency (tf-idf) which represents how important a word is in a document (Rajaraman and Ullman, 2011). However, bag-of-words does not contain any spatial information about the data, i.e., the position of each word relative to others, therefore providing no context. To mitigate this issue ngrams can be used, which are a representation of a contiguous sequence of n words in a given text. For the most used values of n (1, 2 and 3), n-grams are usually referred as unigrams, bigrams and trigrams, respectively. A major drawback of bag-of-words is that the size of feature vector is dependent on the number of distinct words in the dataset and in practice this usually results in large vectors, drastically increasing computation time.



	MARY	IS	HUNGRY	HAPPY	FOR	APPLES	NOT	JOHN	HE	
"Mary is hungry for apples."	1	1	1	0	1	1	0	0	0	[1, 1, 1, 0, 1, 1, 0, 0, 0]
"John is happy he is not hungry for apples."	0	2	1	1	1	1	1	1	1	[0, 2, 1, 1, 1, 1, 1, 1, 1]

Figure 1: Bag of words model (TOPJOBS, 2019)

Word-embeddings: A word embedding is any technique that can directly map a word to a vector of real numbers, often with a much lower dimensionality than bag-of-words vectors. Algorithms using this technique have the capability of learning in which context words are used and thus can capture similarities between similar words. Examples include an embedding layer, used jointly with a neural network and the Word2Vec model (Mikolov et al., 2013). Context is also of extreme importance since it can influence the polarity of key words. The word love, for example, can convey positive emotion in "I love you" but negative emotion in "Real love doesn't exist". Parts-of-speech (PoS) are categories for words with similar grammatical properties such as nouns, verbs and adjectives that are widely used as features to capture context (Go et al., 2009; Kiritchenko et al., 2014). Researchers analyzing the polarity of social media text posts also make use of features prevalent in this type of data such as emoticons, repeated characters (e.g. "I am huuuungry") and punctuation ("You look amazing!!!!") (Kiritchenko et al., 2014).

2.1.2 Lexicon-based approaches

One of the most basic tasks in sentiment analysis is classifying the polarity of a given text, i.e., whether a certain opinion expressed is positive, negative or neutral. Lexicon based approaches are perhaps the most simple way to determine polarity. They consist of assigning a score to each word present in a certain dictionary to represent how positive/negative it is perceived. Typically, the scoring scale goes from -1 (extremely negative) to 1 (extremely positive) with 0 being completely neutral, although some authors consider different scales.

In order to have good results with this approach it is necessary to assign correct scores to each word. To do this, researchers usually resort to crowdfunding by having multiple people give their own opinion on the polarity of words. A well-known algorithm which uses this approach is SentiStrength (Thelwall et al., 2010). The authors collected comments of Myspace and asked 3 oracles to rate the words on a scale of 1 to 5 for both positive and negative emotion, with 1 meaning no emotion and 5 meaning a very strong emotion resulting in a lexicon with 298 positive terms and 465 negative terms with at least a score of 2. Other factors such as repeated punctuation, emoticons and modifier words were also used for changing the score of a particular word in order to improve classification accuracy.

One of the major flaws with SentiStrength is that the score of the words present in the lexicon is fixed and context independent. To solve this problem Saif et al (Saif et al., 2016) proposed "SentiCircles". A "SentiCircle" is a circular representation which, for a term m, contains a list of terms which appear frequently with m and how they influence the polarity of the main term. The calculation of a "SentiCircle" is done by using term correlation and prior sentiment. Term correlation is computed in a similar manner to the tf-idf statistic and indicates the degree of influence (Stronger correlation means more influence) while the prior sentiment is based on external sentiment lexicons (MPQA (Deng and Wiebe, 2015), SentiWordNet (Esuli and Sebastiani, 2006) and SentiStrength (Thelwall et al., 2010)) and specifies how a context term influences the main term (i.e., in a positive or negative manner).

Despite lexicon-based approaches being straight-forward, they often fail to obtain very good accuracy since, like mentioned previously, the polarity of words is extremely context dependent and therefore relying on just word scores can mislead a classifier.

2.1.3 Machine learning approaches

Machine learning is extremely popular for many tasks, including sentiment analysis. In this approach the dataset is divided into two sets: the training set and the testing set. The training set is used for training a model that can optimize a decision function based on the patterns in the data while the testing set is used for making predictions and evaluation, since it contains data that has not been seen previously by the model.

For tasks such as classification and regression, the many approaches can be divided into 3 main categories: supervised learning, unsupervised learning and semi-supervised learning. The difference between these methods is the amount of data labeling required. Supervised learning requires labels for all the samples in the data while unsupervised learning requires no labels at all. Semi-supervised learning sits in the middle requiring labels for only a portion of the data.

By having all the data labeled (as in the case of supervised learning) it is easy to evaluate a model since its predictions can be directly compared to the ground truth. However, labeling data is often an expensive and difficult process (Roh et al., 2019). In some cases, a high level of expertise is needed (e.g., detection tumors in medical images) while in others, such as sentiment analysis, the subjectivity of the task requires multiple labelers to eliminate bias and poor-quality samples. Therefore, an alternative is unsupervised learning as it only requires the data. These types of algorithms can discover the natural groups in data. Nevertheless, model evaluation becomes difficult since there is now nothing to compare the predictions to. To mitigate the disadvantages of the two previous approaches semi-supervised learning can be used. Semi-supervised learning combines a small amount of labeled data (improving accuracy compared to unsupervised learning) with a large amount of unlabeled data (avoiding labeling costs).

In supervised learning, a common model used is the Support Vector Machine (SVM), shown in Figure 2 (Go et al., 2009; Kiritchenko et al., 2014). It has been widely applied to many classification and regression tasks since it shows very good performance. It takes as input data in an n-dimensional space and finds the optimal hyperplane that divides the data into two different classes by maximizing the margin between classes, i.e., maximizing the distance between the hyperplane and the closest point(s) of each class. In multiclass classification (3 or more classes) there are two strategies used: one-vs-one and one-vs-all. In one-vs-one a model is trained for each unique pair of classes and the class is chosen by a majority voting of all trained models while in one-vs-all, for each unique class, a classifier predicts the class confidence and the class with maximum confidence is chosen.

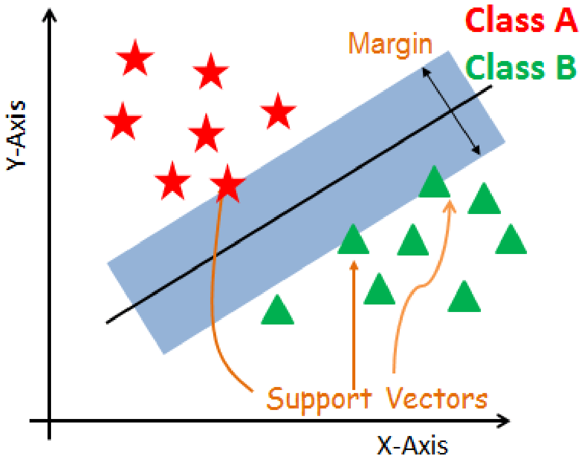


Figure 2: SVM model (Navlani, 2019)

While SVM is a very robust classifier it has the labeling problem of supervised learning. To solve this Go. Et al (Go et al., 2009) built a dataset by querying the Twitter API for emoticons and automatically labeled the data based on this. Emoticons such as ":) and ":D" are assumed to be positive while ":(" is assumed to be negative. A small subset was then manually annotated to improve label reliability reducing the dataset to 177 positive and 182 negative tweets. Features considered included unigrams, bigrams, unigrams and bigrams combined and parts of speech. For classification, the models used were SVM and Naïve Bayes (model based on Bayes' probability theorem) and the best accuracy was 83%. Another option is the use of multiple classifiers, also known as an ensemble classifier. Da Silva et al. (Da Silva et al., 2014) used 5 different classifiers to perform sentiment analysis on multiple datasets present in the literature. To obtain a final prediction two method were used: average of all individual classifiers' predictions and majority voting. For 4 out of the 5 datasets the authors managed to obtain better results than the rest of the literature demonstrating that ensemble classifiers can be a good alternative.

Artificial neural networks have also been used for sentiment analysis. These are comprised of multiple layers of "neurons", including an input and an output layer. Each neuron has a weight associated with it that determines how much the neuron is "activated". When training these weights are updated in an effort to find the optimal weight combination for maximum accuracy. Tang et al (Tang et al., 2014) extracted multiple ngrams features (unigram, unigrams+bigrams and unigrams+bigrams+trigrams) from a Twitter dataset and then trained 2 simple neural networks to learning embeddings of each ngram. Then 3 convolutional layers are used to extract a representation of a tweet based on the learned embeddings.

Neural networks can also be used for extracting features in the form of word embeddings. A particular type of neural network architecture which is well suited for sequential data are recurrent neural networks (RNNs). These networks can remember what they learned before by keeping a hidden state which is a representation of previous inputs. However, RNNs can suffer from a problem called "vanishing gradient", especially when dealing with long sequences. The gradient is used to control by how much weights are updated, and for networks with many layers, this can result in extremely big or extremely small changes, destabilizing the network. Long short-term Memory networks (LSTM) are a type of RNN which address this problem and therefore can capture longer dependencies in sequential data. LSTMs were used by Vadicamo et al (Vadicamo et al., 2017) to extract features from tweets which were used by an SVM for classification.

2.1.4 Hybrid approaches

Hybrid approaches combine the lexicon-based and machine learning methods. The main idea is using previous knowledge from lexicons to obtain additional features from the data to improve a classifiers' performance. Using data from Twitter, Zhang et al (Zhang et al., 2011) used a word polarity lexicon to identify opinionated tweets and label them automatically. Then, the researchers used a SVM for classification. To solve the problem of misassigned labels, the testing set was composed of manually labeled tweets.

Since most available lexicons are for a general purpose, Ghiassi et al (Ghiassi et al., 2013) presented a different approach. To improve the classification accuracy, the authors built a lexicon derived from the same dataset used by the classifiers, extracting features such as words and emoticons that appear frequently on a specific polarity. These features were then used to train a SVM and a neural network obtaining good results.

2.1.5 Discussion

The main advantage of lexicon-based approaches is the fact that they are based on human knowledge and therefore can function as a good starting point for obtaining good results. Nonetheless, calculating sentiment is rule-based and the creation of good rules is difficult since there many contexts to consider and a sentence can be written in many ways. Machine learning classifiers are better equipped to deal with this because they are capable of learning. This is especially useful in domains such as social media where polarity indicators like abbreviations are common and not always present in lexicons. However, learning from the data means that classifiers are domain dependent. As mentioned previously another problem is the amount of labeled data. Available datasets with good quality labels are small and performance is usually worse with smaller datasets. By mixing both approaches, hybrid methods can overcome the problems of individual approaches.

2.2 Image analysis

2.2.1 Feature extraction

In image processing, traditional methods are based on the extraction of low-level features. These can be divided into two categories: color and texture. For color-based features, simple statistical methods such as color histograms can be used. However, in the case of texture-based features, these are usually calculated with more complex algorithms like SIFT-descriptors (Lowe, 1999) and gabor filters (I and Sagi). Nevertheless, low-level features are not very interpretable. Therefore, for sentiment analysis, researchers use lower-level features to derive mid-level features.
Another popular approach is the use of deep learning, more specifically convolutional neural networks. This type of neural network is well suited for image processing and can automatically learn features from images, minimizing manual work.

2.2.2 Mid-level representations

To represent mid-level features, Yuan et al (Yuan et al., 2013) defined 102 attributes based on ease of detection and interpretability, such as different materials (e.g. metal and vegetation) and activities (e.g. playing, cooking). The detection of these attributes was done by training a machine learning

classifier on low-level features.
Another interesting approach is the use of adjective-noun pairs (ANPs) to describe an image (e.g., cute flower, smiling face) by Berth et al (Borth et al., 2013). Considering 1200 ANPs, they trained a model called SentiBank that determines how accurately each ANP describes an image. Sentibank's output can then be used as features for sentiment prediction.
Given that objects often play an important role in conveying emotion, Chen et al (Chen et al., 2014b) made use of object detection models to locate objects in images. Then, they extracted several low-level features from the object as well as the background and used these together with ANPs to improve performance.
In previous approaches ANPs are exclusively used as mid-level features. However, they also carry textual sentiment value. Therefore, Li et al (Li et al., 2018) proposed a method which combines ANPs as mid-level features with their sentiment value.

2.2.3 Deep learning approaches

Deep learning involves the use of models with many layers. It is currently one of the most used approaches for many fields as it has shown good performance. Convolutional neural networks (CNNs) are the go-to model for image preprocessing. These networks are based on convolutions which are the process where a filter "slides over" an image. Each layer of the network contains many filters which can be combined to represent many different set of features.
Since deep learning neural networks take a long time to train from scratch, researchers often use transfer learning to save time. Transfer learning is the process of taking a pre-trained model and retraining on a new, similar task. Earlier layers of a network learn simpler and more generic features, while deeper layers learn more complex features. Retraining is done by "freezing" the weights of the earlier layers, i.e., not changing them, meaning only the weights of the deeper layers will be updated.
Besides the training time, good quality data is also an issue. Some datasets that have high quality labels only have at most a few thousand samples, which is not enough for a model to learn from, while other with a bigger sample size are mostly unlabeled.
Vadicamo et al (Vadicamo et al., 2017) tackled this by using noisy labeling on a large-scale Twitter image dataset. In Twitter, images are typically

accompanied by a description. Working on the assumption that the sentiment of both components is correlated, the authors trained a classifier on the descriptions and then used its' predictions to label the images. To classify these, they used transfer learning with widely used models such as VGG19 and HybridNet.

Not only that, but images in the dataset are also often noisy, i.e., hard to classify even for a person. To solve this, You et al (You et al., 2015) used a Progressive CNN (PCNN). In this approach, some of the samples are removed if the model is too uncertain about them. The new subset is then used to restart the training process and the removal process is repeated. By removing low-quality data during the training process, performance can be improved.

Besides direct classification, deep learning can also be used in other ways. Many studies (Chen et al., 2014a; Jou et al., 2015; Ahsan et al., 2017) have trained CNNs to classify images as ANPs.

Another approach based on object detection (Sun et al., 2016) can be seen in figure 3. First, the authors train a model for sentiment classification. Then, for every image, the authors use R-CNN to extract regions with objects. These regions are evaluated by the trained model, which assigns a score to each region. The best scores are then used together with the whole image score for predicting the sentiment.

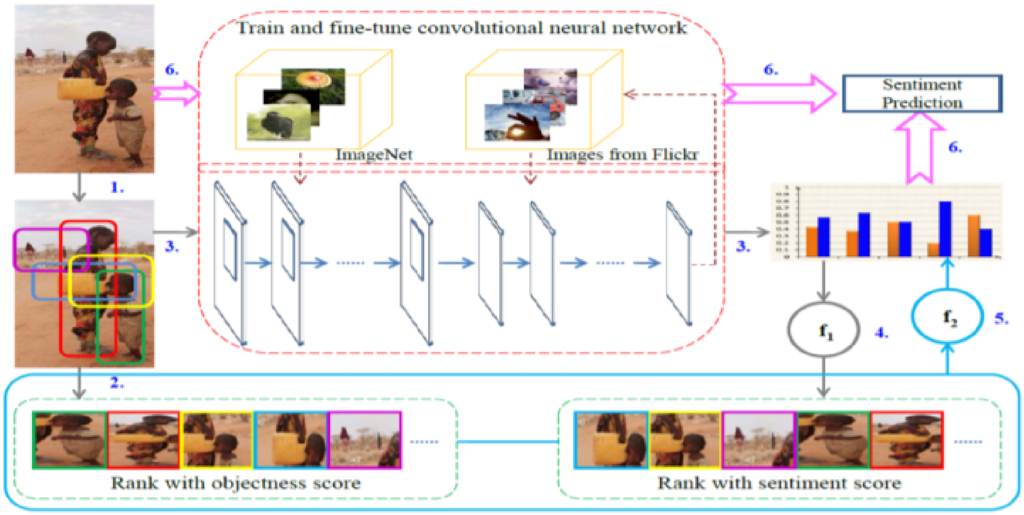

Train and fine-tune convolutional neural network
ImageNet
Images from Flickr
Sentiment Prediction
1.
2.
3.
3.
4.
5.
6.
6.
6.
f_1
f_2
Rank with objectness score
Rank with sentiment score

Figure 3: Architecture of the approach used in (Sun et al., 2016)

VGG

VGG models were proposed by the Visual Geometry Group (Simonyan and Zisserman, 2015). While previously proposed multi-layer convolutional neural network architectures had variable filter sizes for each layer, VGG models have one unique filter size for every layer in the network. Because of a reduced filter size, the models have an increased number of layers. The authors assume that several smaller filters can work as well as fewer, bigger filters. These models have shown very good performance, but the high number of layers can slow down training.

Residual network

Residual neural networks were first proposed in (He et al., 2015) as a way to solve the vanishing gradient problem. In very deep neural networks based on gradient learning methods, weights in early layers may stop updating as the gradient becomes too small to make a significant change, effectively stopping the training process. To overcome this, resnets implement shortcut connections (connections that skip layers) and identity mappings. Identity mappings ensure a similar input and output causing the gradient to remain stable. Without vanishing gradients it is possible to increase the depth of neural networks which may increase performance.

YOLO

YOLO is an object detection model first proposed in (Redmon et al., 2015). To detect objects, this model first divides an image into square grids. Then, for each grid a predefined amount of bounding boxes is computed, as well as their confidence and a class probability map. The confidence of bounding boxes is multiplied with the class probability map to obtain the final detections. YOLO's object detection process is shown in figure 4.

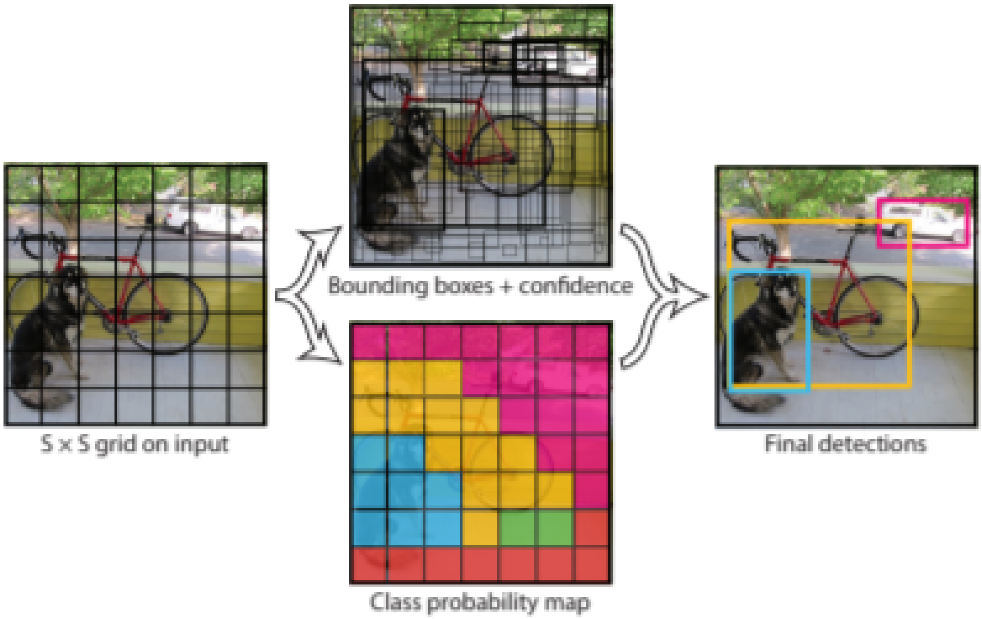


Figure 4: YOLO object detection process (Redmon et al., 2015)

2.2.4 Discussion

Mid-level features can be used as a good representation of an image. When used together with object detection models, they can achieve even better performance as objects usually have a great impact on visual sentiment. Deep learning approaches translate an image directly into sentiment. This can hinder performance as it is a different process from human perception. The main problem for the analysis of visual content is data unreliability. Because large datasets are mostly unlabeled, researchers rely on the text associated with the data as pseudo-labels. These can pose problems as they are not always accurate.

2.3 Multimodal analysis

2.3.1 Combining multiple modalities

There are many different strategies for joining multiple modalities, such as early fusion, late fusion and attention. In early fusion, features are extracted from each type of data and are then combining into a single feature vector as input for one classifier. On the other hand, in late fusion, each modality is first classified and then all the scores are used together for a final prediction. Attention (Vaswani et al., 2017) is a more recent technique which has been gathering popularity ever since its introduction by Vaswani and colleagues. It works by calculating importance weights of each individual feature freeing models to focus on the most relevant parts of an input. Figure 5 demonstrates how attention highlights certain features of the input.

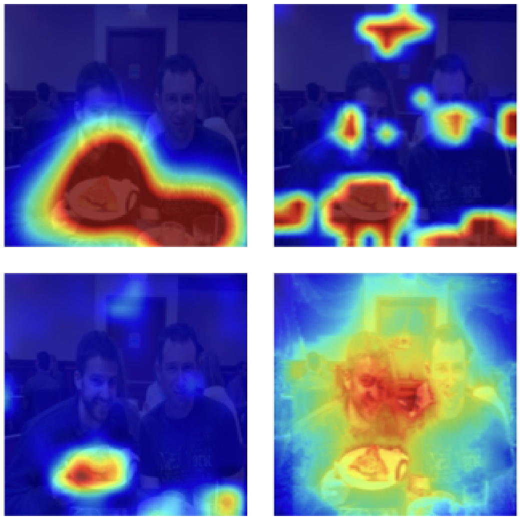


Figure 5: Attention visualized (Das et al., 2016)

2.3.2 Deep learning approaches

For multimodal data, sentiment analysis is mostly based around the use of deep neural networks. The main difference between the many approaches is the technique used.

Cai et al (Cai and Xia, 2015) trained two CNNs, one for text and one for images, as shown in Figure 6. Then, they extracted the features from the penultimate layer of each network to feed to a new CNN which is then used for classifying the sentiment. This approach may allow a neural network to learn from multiple sources at once.

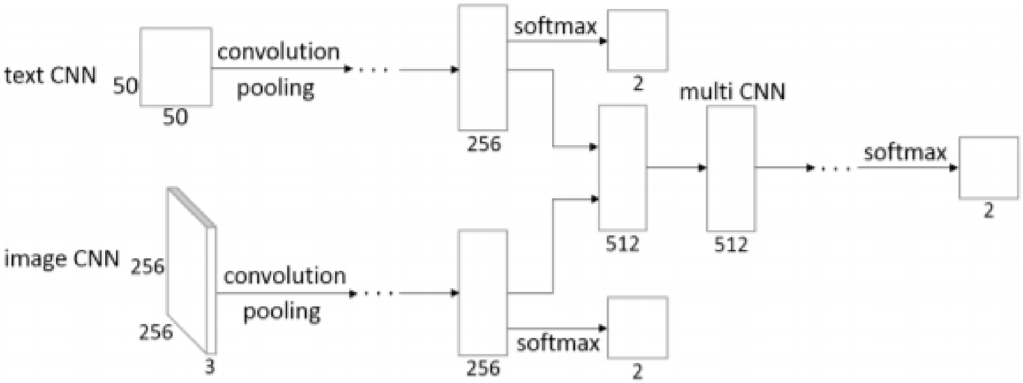


Figure 6: Feature-level fusion approach used in (Cai and Xia, 2015)

Another approach based on late fusion is made in (Gaspar and Alexandre, 2019). Using a textual and visual content dataset, the authors used different machine learning models to predict text and image polarity. Furthermore, they used a statistical method to determine the probability of an object belonging to each class. Finally, the results of each method were joined together via weighted voting. As deep learning models do not have very good performance on large scale datasets using other features such as visual content may improve results.

Some researchers have also analyzed the correlation between different modalities. Chen et al. (Chen et al., 2017) used weighted co-training of two different networks, one for textual and another for visual data. Two datasets are considered, one labeled and one unlabeled. Weighted co-training is an iterative process. It starts by training the two networks on the labeled set, followed by calculating the similarity of text and images belonging to each unlabeled instance. Then, the most confident ones are moved to the labeled set and the whole process is repeated.

2.3.3 Attention-based approaches

Xu et al. proposed a co-memory attention network (Xu et al., 2018). This model consists of two separate networks with many ”hops” where information is continuously shared between networks. This process works by feeding visual features from the previous hop to the textual network and textual features to the visual network. The architecture of a co-memory can be seen in Figure 7.

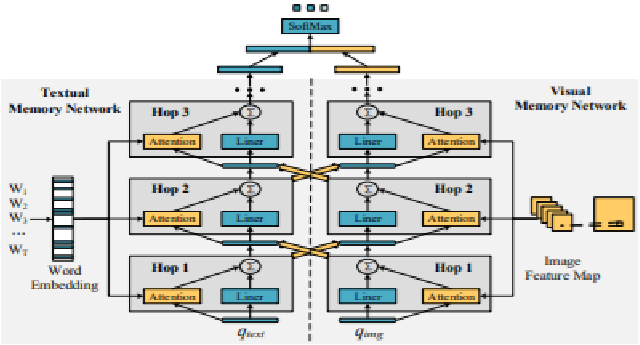


Figure 7: Co-memory network architecture (Xu et al., 2018)

Xu et al. presented a Hierarchical deep fusion model (Xu et al., 2019b) where they exploited not only textual and visual features but also the links between social data. Examples of these links include comments by the same user, common image tags and submissions to the same social media groups. Working with the CMU-MOSI dataset (Zadeh et al., 2016), Chen et al. introduced Gated Multimodal Embedding LSTM (Chen et al., 2018). CMU-MOSI is composed of annotated youtube videos and therefore is it crucial to understand which moments are most important for sentiment analysis. Thus, in Chen et al.'s approach, attention is calculated at each time step. Another important feature is the presence of on/off gates which control whether a particular time step contributes to the final result. They managed to obtain an improvement of 3% in accuracy and 1.1% in F-score.

Xu et al. (Xu et al., 2019a) made use of attention mechanisms in a bi-directional manner. The authors presented bi-directional multi-level attention networks which are able to take advantage of how text and images influence each other. With this approach, the model is able to make predictions with over 85% accuracy on multiple datasets.

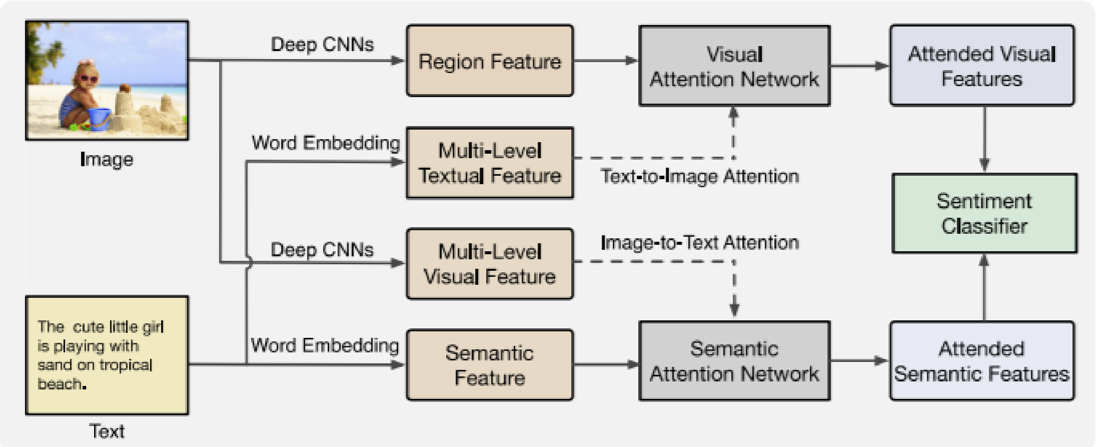


Figure 8: Framework of the bi-directional attention network (Xu et al., 2019a)

2.3.4 Discussion

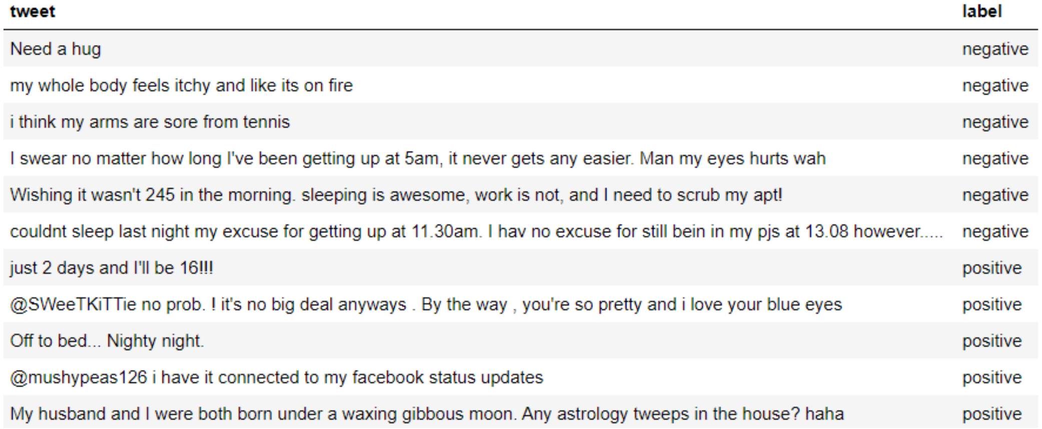
The main difficulty in multimodal approaches is understand how to combine multiple types of data. Late fusion is a common strategy as it very simple and easy to implement. The main advantage of this method is that classifiers are trained on a single modality, meaning they are well suited to understand the features of a specific modality. However, given that early fusion joins multiple sets of features, the correlation of different modalities can be better captured by a model. To make use of the advantages of both strategies, hybrid fusion has also been used (Nemati and Naghsh-Nilchi, 2017). This process consists of performing late fusion on an early fusion model and all the models trained on each modality.
Attention has been getting more popularity in multimodal sentiment analysis. A key advantage of this approach is that it is able to take advantage of the correlation of multiple modalities better than early fusion and can guide the models towards the most important features.

2.4 Datasets

2.4.1 Sentiment140

Sentiment140 is a textual dataset used in (Go et al., 2009) composed of 1.6 million tweets. The researchers queried the Twitter API with emoticons to automatically label each sample as either positive (for emoticons such as

":)" and similar) or negative (":(" and similar). The dataset contains 800,000 positive and negative tweets. Figure 9 contains some examples of tweets and labels present in sentiment140.

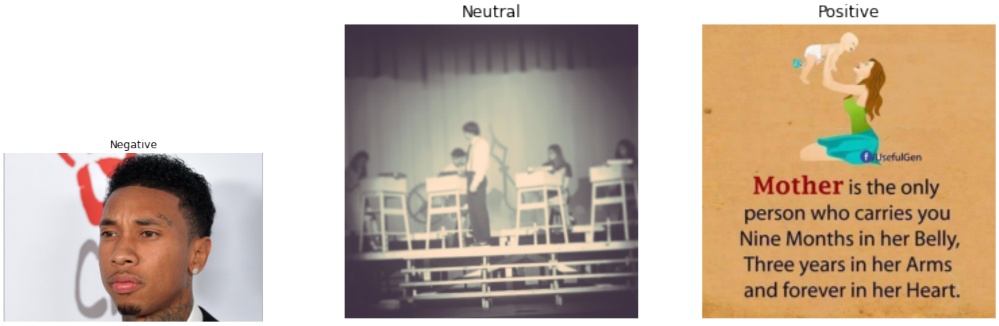


tweet	label
Need a hug	negative
my whole body feels itchy and like its on fire	negative
i think my arms are sore from tennis	negative
I swear no matter how long I've been getting up at 5am, it never gets any easier. Man my eyes hurts wah	negative
Wishing it wasn't 245 in the morning. sleeping is awesome, work is not, and I need to scrub my apt!	negative
couldnt sleep last night my excuse for getting up at 11.30am. I hav no excuse for still bein in my pjs at 13.08 however.....	negative
just 2 days and I'll be 16!!!	positive
@SWeeTKiTTie no prob. ! it's no big deal anyways . By the way , you're so pretty and i love your blue eyes	positive
Off to bed... Nighty night.	positive
@mushypeas126 i have it connected to my facebook status updates	positive
My husband and I were both born under a waxing gibbous moon. Any astrology tweeps in the house? haha	positive

Figure 9: Example of tweets in the dataset sentiment140

2.4.2 T4SA

T4SA is a dataset used by Vadicamo et al. (Vadicamo et al., 2017). They collected 3.4 million tweets that contain both text and images. Then, the tweets with the most confident sentiment were selected resulting in 1,179,957 tweets and 1,473,394 images (note: a tweet may contain multiple images associated). To train their models, corrupted and near-duplicate images were removed and the dataset was balanced. This new subset was named B-T4SA and contains an equal amount of positive, negative and neutral images (156,862) for a total of 470,586 images, as well as the associated text with 384527 samples. The researchers also provided the division used for the training, validation and testing subsets. Figure 10 shows an example of negative, neutral and positive tweets in the dataset.


Negative
Neutral
Positive
UsefulGen
Mother is the only person who carries you Nine Months in her Belly, Three years in her Arms and forever in her Heart.

(a) **Negative:** Tyga received an unwanted gift for his birthday - a new lawsuit.

(b) **Neutral:** Here's a sneak peak of the unfinished set of These Shinning Lives #prospertheatre2k16 #Theseshinninglives

(c) **Positive:** Of all the gifts that life has to offer, a loving mother is the greatest of them all.

Figure 10: Examples of images and respective tweets in the T4SA dataset

2.4.3 Flickr dataset

Katsurai and Satoh (Katsurai and Satoh, 2016) published a dataset based on flickr. This websites is photo-based, meaning it is well suited for sentiment analysis. In addition to images, flickr posts also have a title, a description, tags and other elements.

The dataset contains the id of the images as well as the opinion of 3 annotators (positive, negative or neutral). In total there are over 90 thousand samples. The researchers used the images presented and textual features to analyse multimodal sentiment. However, there is not much detail in regards to which features were considered for analysing text. Therefore, it is hard to compare results with the ones present in the paper.

2.4.4 Multi-view Sentiment Analysis (MVSA)

MVSA (Niu et al., 2016) is a multimodal dataset composed of images and text extracted from Twitter. There are two variants of the dataset, named MVSA-Single and MVSA-Multiple. MVSA-Single is labelled by one annotator and MVSA-Multiple is labelled by three. One of the defining features of this dataset is that labels are independent, i.e, one person can have a different opinion on the polarity of the text and image that make up a tweet. The distribution of annotations in MVSA-Single and MVSA-Multiple can be seen in tables 1 and 2, respectively.

	Positive	Neutral	Negative	Total
Text	1731	1921	1217	4869
Images	2708	938	1223	4869

Table 1: MVSA-Single annotations

	Type of data	Positive	Neutral	Negative	Total
Annotator 1	Text	9724	8072	1799	19595
	Image	9799	7423	2373	19595
Annotator 2	Text	10205	7605	1785	19595
	Image	9796	7475	2324	19595
Annotator 3	Text	9366	8816	1413	19595
	Image	9211	8569	1815	19595

Table 2: Distribution of classes per annotators and type of data in the MVSA-Multiple dataset

2.5 Conclusion

Traditional approaches were used in the early days of sentiment analysis. However, due to their simplicity, results were lackluster. Nowadays, with the advancement in computing power and research, machine learning approaches have shown better capability. Deep learning in particular is very popular in the literature due to its high performance. Attention is a recent technique applied to deep learning models that improves results even further.

There are many datasets available for sentiment analysis. Sentiment140 and T4SA are datasets for text and images, respectively, with a very large sample size. However, they are annotated automatically, which might compromise their quality. MVSA and flickr (Katsurai and Satoh, 2016) are multimodal datasets with strong labelling. Nevertheless, their lower sample size may hinder the learning capabilities of machine learning models.

3 Proposed approach

The first chapter introduced sentiment analysis and detailed its uses, such as assessing public opinion about public figures, brands and present events. The main challenges, including lack of research, modality fusion and label reliability were also discussed.
Next, in the state of the art, methods used in the literature such as traditional, deep learning and attention-based approaches were presented. Following this, popular datasets were described.
In this chapter, the methods used in this work will be explained, including data preprocessing, models used and how the results are obtained and evaluated.

3.1 Datasets used

In the early stages of this work MVSA was not available. Therefore, many datasets were used, including sentiment140, flickr and T4SA. However, each of these datasets had its flaws. sentiment140's tweets were classified as either positive or negative based on the queried emoji. T4SA's images were labelled by a model's prediction of the text that accompanied the image. In both of these cases, the annotation process was not performed by humans which can lead to weak labelling. Flickr posts contain a photograph with a title and an option to add a description and tags. Nevertheless, Katsurai and Satoh only provided the id of the post and its label. Because of this, all the data must be downloaded from flickr and some samples are missing. Another issue is the lack of clarity in regards to which parts (title, description, tags, etc.) of a post were used as the textual modality of the data.
MVSA-Multiple was the main dataset worked on as it had strong labelling (with multiple annotators) as well as a good sample size (19595).

3.2 Data preprocessing

Since the data consists of both text and images, different preprocessing techniques are required for each case.
Images come in various sizes and need to be resized a standard size. In some cases, the objects present in images were also extracted. Image augmentation techniques such as rotation, zoom and shifting pixels can also be used to increase the amount of data available.

Data type	Voting agreement	Positive	Neutral	Negative	Total
Text	Majority (2+ votes)	10058	7285	1063	18406
	Unanimous (3 votes)	3806	2580	241	6627
Image	Majority (2+ votes)	9933	6759	1468	18160
	Unanimous (3 votes)	3785	2395	351	6531

Table 3: Agreement between annotators in the MVSA-Multiple dataset

To process text, there are many techniques to be considered. The first step is to remove noise, including emails, URLs, punctuation (excluding emojis), numbers and keywords related to the dataset (in the case of Twitter: RT for retweet and @username for usernames). Next, PoS (parts-of-speech) can be extracted. Afterwards, all uppercase letters are converted to lowercase and then stop words (eg. that, this) are removed. Finally, lemmatization can be applied. To transform the data into numeric form, GloVe embeddings (Pennington et al., 2014) were used.

3.3 Handling multiple annotators

MVSA-Multiple has 3 annotators. Because of this, differences of opinion can manifest as can be seen in table 3. To solve this problem, the final label is determined by the majority, i.e., if an image has at least 2 positives it is considered positive. Cases where there is no clear agreement, i.e., only 1 vote per class, are disregarded.

Given the subjective nature of sentiment analysis, another factor that may influence results is the bias of annotators. Thus, tests were also performed on each annotator to understand how the model adapts to each one and how the results differ.

3.4 Choosing models and parameters

The first step to working with data is choosing a model. For text, traditional machine learning models such as Support Vector Machines (SVMs) were common. Recently, deep learning models (for example LSTM) have been rising in popularity. For images, convolutional neural networks have been almost exclusively used for classification tasks since AlexNet (cite alexnet). The models tested in this work were LSTMs for text and various CNN architectures available in the keras framework (VGG19, InceptionV3, ResNet variations) for images. When carrying out multimodal tests, the best performing models of each modality were used with early fusion and attention. Deep learning models were chosen for a couple of reasons. Firstly, these are widely used in the literature with good results. They are also able to be combined for analysing multimodal data when using attention-based methods and early fusion due to features from each modality being able to be easily extracted and combined.

3.5 Sampling strategies

As discussed previously, large class imbalances can lead to problems when training machine learning models. To resolve this problem, manipulating the number of samples can be considered. Two common methods employed for this purpose are undersampling and oversampling. Undersampling consists of removing examples from the majority class while oversampling introduces new samples to the minority class. New examples can be obtained by copying already existing ones. However, this method does not add new information. Other approaches such as SMOTE (Bowyer et al., 2011) make small changes to already existing data points to address this issue.

3.6 Metrics

Accuracy measures the percentage of examples correctly classified. It is commonly used for evaluating results as it is very straightforward. However, since only true positives and true negatives are considered, this metric can be misleading. When dealing with a large class imbalance (as in the MVSA dataset), models often gravitate towards the majority class at the detriment of the others. Because of this, models can have high accuracy despite bad performance in the minority class. To address this issue, precision, recall and f-score may be used. Precision is the percentage of correct classifications among all positive predictions while recall measures how many positive samples were identified by the model. F-score is calculated from precision and recall. Figure 11 explains how these two metrics are calculated. The advantage of f-score over accuracy is that it considers and penalizes false positives and false negatives. In multiclass scenarios, it can be calculated for each individual class in a one-versus-all scenario, i.e., one class is deemed positive while the others are negative. For this case, a final F-score can be obtained by averaging all scores (F-Macro) or a weighted average by class frequency (F-Micro). this work, the metrics chosen were:

- **Accuracy**, due to its wide use in the literature
- **F-score per class**, to understand the performance of the model for each class
- **F-Macro**, as it gives equal importance to every class

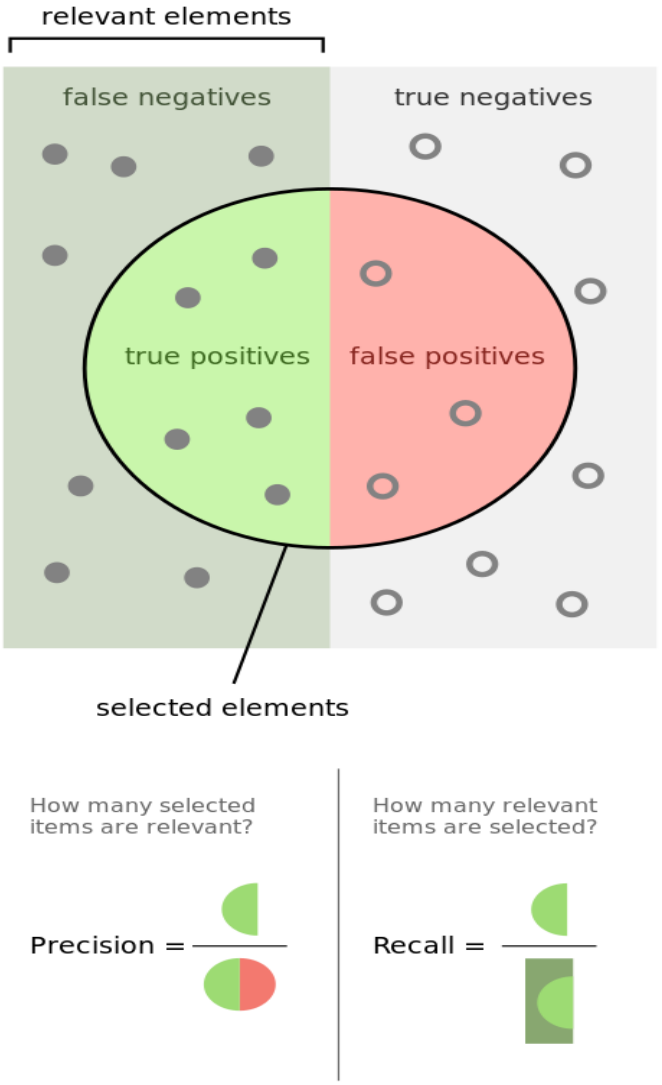


Figure 11: Calculation of precision and recall (Wikipedia, 2014)

3.7 Choosing model parameters

During the training process there are many factors that influence how a machine learning model learns from data. Some of these are optimizers, learning rate and number of epochs.

3.7.1 Optimizers

Optimizers are algorithms that are used to minimize a loss function and determine how a model's weights are updated. Many different versions are available, such as SGD, RMSprop and Adam. Optimizers play a key role in the training process.

3.7.2 Learning rate

The learning rate controls how much the weights are updated throughout training. This hyperparameter is very important as it plays a major role in the final result. If the learning rate is too high, overfitting can happen. Meanwhile, if it is too low, learning is very slow and underfitting occurs. Figure 12 demonstrates these scenarios.

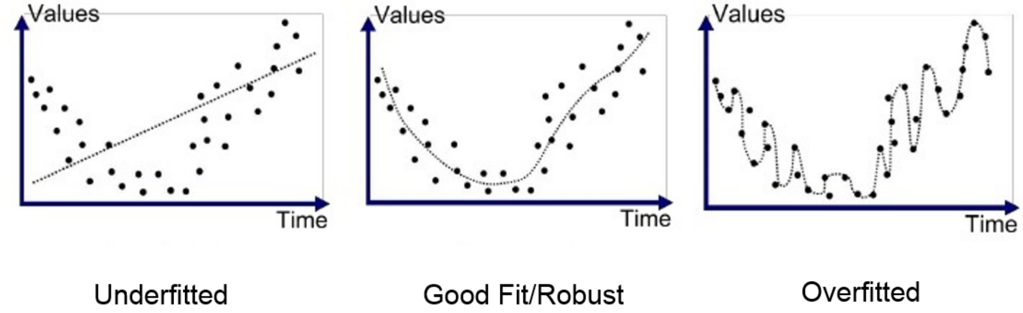


Figure 12: Underfitting, optimal fitting and overfitting visualized (Balaji, 2019)

3.7.3 Epochs

An epoch is a single pass through the entire data. Machine learning models need a lot of data to perform well and when there is not enough, multiple epochs can improve results. However, much like learning rate, too many epochs may lead to overfitting.

4 Results

In this chapter, the results from testing various relevant factors. The first tests were done on individual modalities (i.e. only on text and only on images) to get a better understanding of what works for each modality before studying the multimodal case.

4.1 Optimizer testing

The weights of an artificial neural networks regulate its predictions. Therefore, they play a crucial role in getting good performance. Because of this, it is important to update them correctly during the training process.
The text model tested in this work is the LSTM network. For visual sentiment analysis, many models are considered, including VGG19, Resnet152v2 (He et al., 2015), InceptionV3 (Szegedy et al., 2015) (shown in figure 14) and InceptionResNetV2 (Szegedy et al., 2016) and are all pretrained on the imagenet dataset (Deng et al., 2009). Features were extracted from each network and were processed by a multilayer perceptron with 2 hidden layers of 1024 units each with a ReLu activation function. The optimizers chosen were Adam, SGD and RMSprop and the learning rate was set to 0.0001.

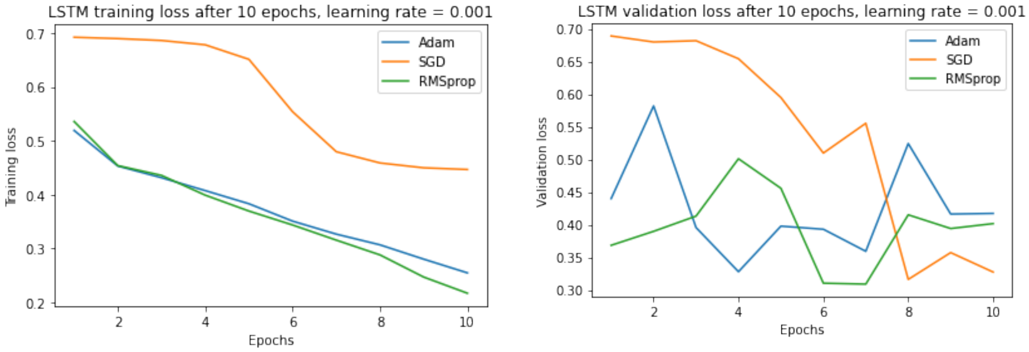


Figure 13: LSTM validation loss for different optimizers (learning rate = 0.001)

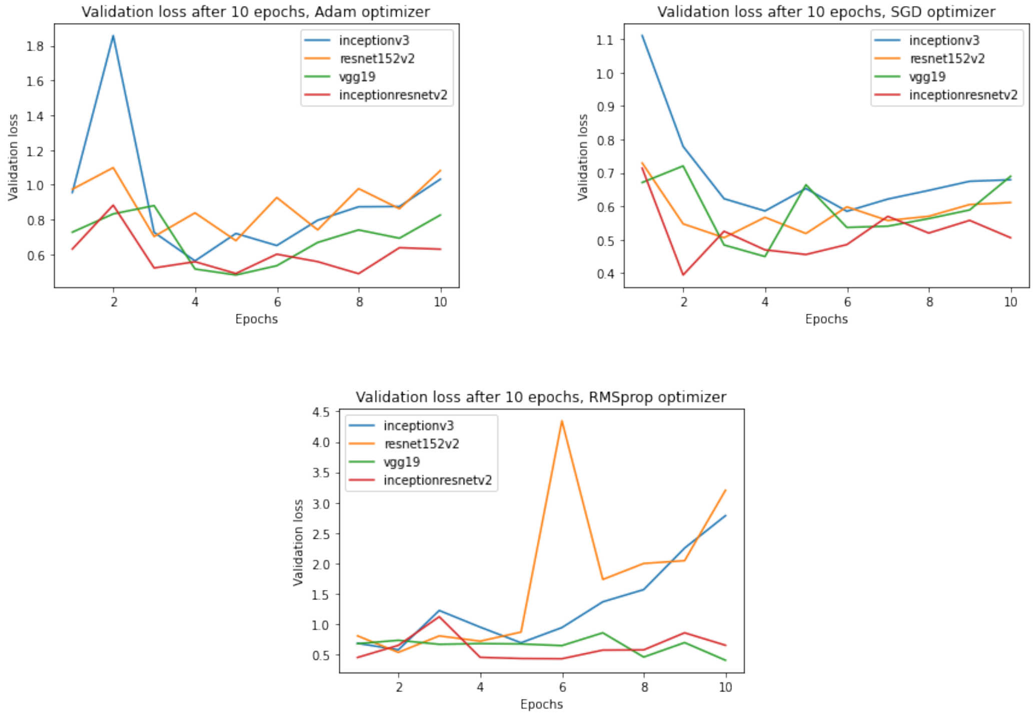


Figure 15: Validation loss of multiple CNNs for different optimizers

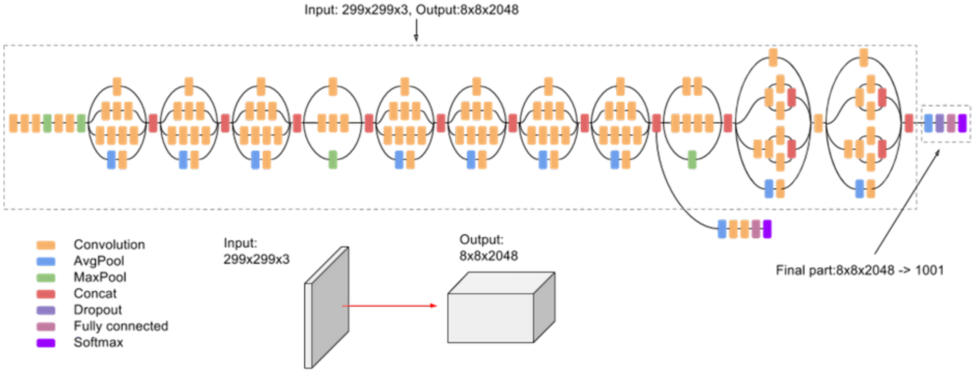


Figure 14: InceptionV3 architecture

Figure 13 presents the test of the LSTM network. Both Adam and RMSprop display good results for this model. Analysing Figure 15, it is clear that InceptionResNetV2 and VGG19 are the best performing models. When it comes to optimizers, SGD and Adam show low values for validation loss.

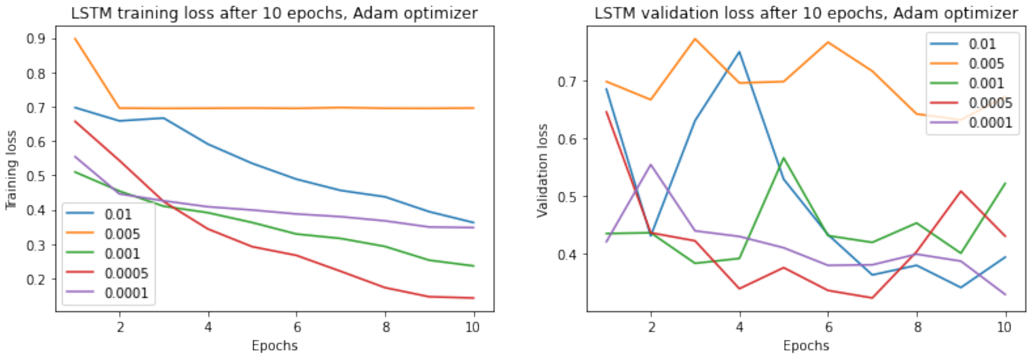


Figure 16: LSTM learning rate tests with Adam optimizer

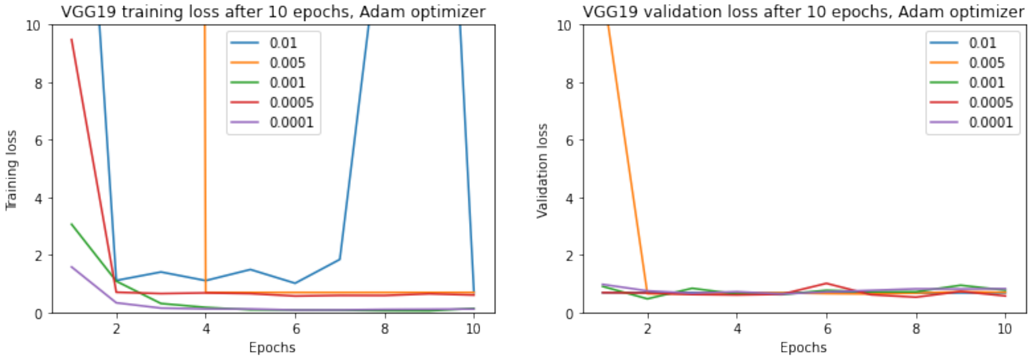


Figure 17: VGG19 learning rate tests with Adam optimizer

For further testing, the image model used will be VGG19. To keep parameters consistent across modalities, the Adam optimizer was chosen since it also performs very well with the LSTM network.

4.2 Learning rate

In order to understand the optimal learning rate many different values were tested (0.01, 0.005, 0.001, 0.0005 and 0.0001). The optimizer chosen was Adam as it showed promising results in both text and image models. The number of epochs was set to just 10 as preliminary tests showed that progress stopped after just a few epochs.
When looking at Figures 16 and 17, 0.001 looks to be the best value for

Method	F-Positive	F-Negative	F-Macro	Accuracy
No method	0.85	0.36	0.60	0.78
Image augmentation	0.87	0.11	0.49	0.77
3 objects	0.91	0.34	**0.62**	0.84
5 objects	0.88	0.37	**0.62**	0.80
7 objects	**0.92**	0.33	**0.62**	**0.86**
10 objects	0.87	**0.38**	**0.62**	0.78

Table 4: Results for different image processing techniques

learning rate as it shows some of the lowest values for validation/training loss in both models. 0.0001 could also be used for the same reasons, although 0.001 was chosen as it converges quicker when looking at LSTM validation loss and at the same speed in training loss.

Past 3 epochs, the text model starts to overfit as validation loss stabilizes while training loss continues to decrease. Meanwhile, the image model shows difficulties in learning as its validation loss does not decrease.

4.3 Image processing methods

Techniques studied for processing images in this work are image augmentation and use of object features. In the latter, objects are sorted by model confidence. Then, features from the top 3, 5, 7, and 10 objects in the image are extracted and concatenated with the features from the whole image.

Looking at table 4, it is clear that image augmentation does not improve results. Despite a slight improvement in the F-score of the positive class, it comes at the detriment of the negative class and thus, also lowering F-Macro. Using object features, results improve slightly. In all cases, F-Positive increases by up to 7%, while F-Macro increases by 2%. With 10 objects, all F-scores are better than the benchmark.

4.4 Individual annotators

In this section, results will be shown for each individual annotator. Given the class imbalance between positive and negative polarity, undersampling and oversampling was also used. Oversampling was done by randomly duplicating data points from the minority (negative) class. Multiple class ratios were also tested. The ratios presented in the tables are always positive:negative. For

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	**0.88**	**0.49**	**0.69**	0.72
Undersampling (1:1)	0.83	0.45	0.64	0.51
Undersampling (2:1)	0.86	0.47	0.67	0.64
Undersampling (3:1)	0.82	0.45	0.64	0.73
Oversampling (1:1)	**0.88**	0.43	0.66	0.78
Oversampling (2:1)	0.86	0.44	0.65	**0.80**
Oversampling (3:1)	0.84	0.44	0.64	0.78

Table 5: Results for text analysis of annotator 1

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	**0.84**	**0.43**	**0.64**	**0.75**
Undersampling (1:1)	0.81	0.41	0.61	0.72
Undersampling (2:1)	0.81	0.41	0.61	0.72
Undersampling (3:1)	0.79	0.40	0.59	0.69
Oversampling (1:1)	**0.84**	0.41	0.62	**0.75**
Oversampling (2:1)	0.83	0.41	0.62	0.73
Oversampling (3:1)	0.79	0.40	0.59	0.69

Table 6: Results for text analysis of annotator 2

example, undersampling (2:1) means that there are 2 positive samples for each negative sample.

4.4.1 Text results

Observing tables 5, 6, and 7 there is clearly an impact on how well the model can adapt to each annotator. Considering the positive class, annotator 3 had the best results with values close to 0.9 in multiple cases, followed by annotator 1 with annotator 2. When it comes to the negative class, annotator 1 is the best. Overall, annotator 1 performs the best with a F-Macro of 0.69. Interestingly, different sampling strategies only improve performance for the third annotator.

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.82	0.40	0.61	0.72
Undersampling (1:1)	0.80	0.38	0.59	0.69
Undersampling (2:1)	0.87	**0.44**	**0.66**	0.79
Undersampling (3:1)	0.88	0.43	**0.66**	0.80
Oversampling (1:1)	**0.90**	0.42	**0.66**	**0.83**
Oversampling (2:1)	0.89	0.42	**0.66**	0.81
Oversampling (3:1)	0.88	0.42	0.65	0.80

Table 7: Results for text analysis of annotator 3

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.86	0.33	**0.59**	0.77
Undersampling (1:1)	0.62	**0.36**	0.49	0.52
Undersampling (2:1)	0.76	**0.36**	0.56	0.65
Undersampling (3:1)	0.82	**0.36**	**0.59**	0.72
Oversampling (1:1)	0.86	0.27	0.56	0.77
Oversampling (2:1)	**0.88**	0.25	0.56	**0.79**
Oversampling (3:1)	0.86	0.27	0.56	0.77

Table 8: Results of image analysis of annotator 1

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.82	0.37	0.59	0.72
Undersampling (1:1)	0.60	0.37	0.48	0.51
Undersampling (2:1)	0.75	**0.38**	0.56	0.64
Undersampling (3:1)	0.83	**0.38**	**0.60**	0.73
Oversampling (1:1)	0.87	0.29	0.58	0.78
Oversampling (2:1)	**0.88**	0.28	0.58	**0.80**
Oversampling (3:1)	0.87	0.32	0.59	0.78

Table 9: Results of image analysis of annotator 2

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.84	0.34	**0.59**	0.74
Undersampling (1:1)	0.71	**0.35**	0.53	0.60
Undersampling (2:1)	0.80	**0.35**	0.57	0.69
Undersampling (3:1)	0.87	0.29	0.58	0.78
Oversampling (1:1)	**0.90**	0.19	0.55	**0.81**
Oversampling (2:1)	0.89	0.26	0.57	0.80
Oversampling (3:1)	0.88	0.29	0.58	0.79

Table 10: Results of image analysis of annotator 3

4.4.2 Image results

For images, learning appears to be difficult independently of the annotators given that all of the best F-scores are very similar. Curiously, despite the increase in negative samples, oversampling leads to sharp decreases in the F-negative score. A possible explanation is that because examples are duplicated, the model overfits on the negative examples which leads to worse performance on non-training data.

4.5 Comparisons with other works

In this section, the results obtained will be compared with those presented by Niu et al. (2016) for each individual modality and for the multimodal case.

4.5.1 Text results

When analysing text, the approaches chosen by Niu et al. were TF and TF-IDF. They compared their results with two popular algorithms, Sentiwordnet and Sentistrength.

This work (Table 11) was able to get good results when it comes to the positive class with a F-Score in 0.94 in the best case. However, correctly identifying negative examples proved to be more challenging with only a maximum F-score of 0.47. Meanwhile, Niu et al. (Table 12) managed to obtain better results in the negative class with a F-score of 0.64 despite lower performance in the positive class.

Method	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.82	0.40	0.61	0.84
Undersampling (1:1)	0.87	0.41	0.64	0.79
Undersampling (2:1)	0.88	0.42	0.65	0.80
Undersampling (3:1)	0.90	0.45	0.68	0.83
Oversampling (1:1)	**0.94**	**0.47**	**0.70**	**0.89**
Oversampling (2:1)	0.92	**0.47**	**0.70**	0.86
Oversampling (3:1)	0.89	0.42	0.66	0.82

Table 11: Results of text sentiment analysis in this work

Method	F-Positive	F-Negative	F-Macro	Accuracy
SentiWordnet	0.64	0.56	0.60	0.60
SentiStrength	0.63	**0.64**	0.63	0.63
TF	**0.79**	0.57	**0.68**	**0.72**
TF-IDF	0.77	0.54	0.65	0.69

Table 12: Results of text sentiment analysis in (Niu et al., 2016)

4.5.2 Image results

For the visual modality, Niu et al. used many different low and mid-level image features such as color histogram, GIST and Sentibank. They also tested feature fusion. LV-Early and LV-Late combine low-level features while V-Early and V-Late fuse all eight features.

In this work (Table 13) and Niu et al.'s (Table 14), the same tendencies are present. This work shows once again better F-positive scores with a

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.85	**0.36**	0.60	0.78
Undersampling (1:1)	0.76	0.33	0.55	0.66
Undersampling (2:1)	0.89	**0.36**	0.62	0.83
Undersampling (3:1)	0.92	0.32	0.62	0.88
Oversampling (1:1)	**0.93**	0.26	0.59	**0.89**
Oversampling (2:1)	0.92	0.27	0.59	0.88
Oversampling (3:1)	0.91	0.35	**0.63**	0.87

Table 13: Results of image sentiment analysis in this work

Method		F-Positive	F-Negative	Macro-F1	Accuracy
Low-level	Color histogram	0.77	0.26	0.52	0.65
	GIST	0.78	0.15	0.46	0.65
	LBP	0.79	0.07	0.43	0.65
	BoVW	0.77	0.35	0.56	0.66
Middle-Level	Classemes	0.75	**0.43**	0.59	0.65
	Attributes	0.78	0.4	0.59	0.68
	Sentibank	**0.79**	0.38	0.58	**0.69**
Aesthetic	Aesthetic	0.78	0.18	0.49	0.66
Fusion	LV-Early	0.78	0.36	0.57	0.67
	LV-Late	0.79	0.34	0.56	0.68
	V-Early	0.78	0.42	**0.60**	0.68
	V-Late	0.79	0.38	0.58	0.68

Table 14: Results of image sentiment analysis in Niu et al. (2016)

Sampling	F-Positive	F-Negative	F-Macro	Accuracy
No sampling	0.93	0.27	0.60	0.87
Undersampling (1:1)	0.89	0.38	0.64	0.82
Undersampling (2:1)	0.94	**0.44**	**0.69**	0.89
Undersampling (3:1)	0.86	0.32	0.59	0.77
Oversampling (1:1)	**0.96**	0.19	0.57	**0.93**
Oversampling (2:1)	**0.96**	0.28	0.62	0.92
Oversampling (3:1)	**0.96**	0.21	0.58	**0.93**

Table 15: Results of multimodal sentiment analysis in this work

maximum of 0.93 against Niu et al.'s 0.79. Both approaches also demonstrate subpar performance with values of 0.36 and 0.42, respectively.

4.5.3 Multimodal results

Table 16 shows the results of there different models implemented by the MVSA researchers. T-V-Early and T-V-Late implement early and late fusion, respectively, from TF textual features and low and mid-level image features. M-DBM is a multi-model Deep Boltzmann Machine Srivastava and Salakhutdinov (2014). For this work, an early fusion model is used, with features extracted by the LSTM and VGG19 models presented previously.

Method	F-positive	F-negative	F-macro	Accuracy
T-V-Early	**0.82**	**0.57**	**0.7**	**0.75**
T-V-Late	0.81	0.45	0.64	0.73
M-DBM	0.82	0.53	0.68	0.75

Table 16: Results of multimodal sentiment analysis in Niu et al. (2016)

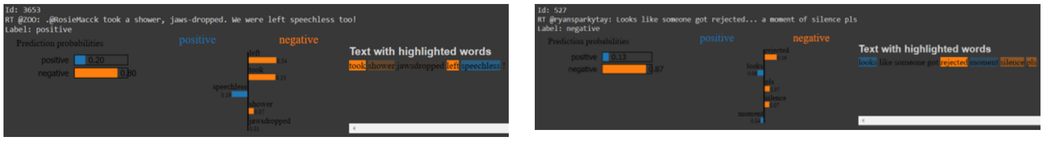
Much like images, this work's multimodal results (Table 15) show very high F-positive scores with values over 0.9 in almost all cases while F-negative only reaches a maximum of 0.44. Oversampling also shows the same tendency of decreasing performance in the negative class. The models studied by Niu et al. showed a more balanced classification, with better F-negative scores.

4.5.4 Discussion

Overall, the performance of the models studied in this work suffers due to the class imbalance present leading to an over-representation of the positive class in the tests performed. Niu et al. also show the same problem, although to a lesser degree. Both approaches demonstrate difficulty in learning to classify images. One possible explanation is that there is no clear defining features on what makes a certain image be considered positive or negative. Because of this, there may be a need for a lot more data in order for a model to understand the polarity of visual data. This issue also appears in multimodal models as they are partially trained on images. However, Niu et al. are better able to take advantage of the interactions between both modalities as their results are more balanced.

4.6 Analysing model's predictions

Deep learning models are often called black boxes as it is not possible to understand the factors that influence their predictions. This can lead to a lack of trust in the results provided as it is unclear what is the reason behind a model's decisions. Addressing this, Ribeiro et al. created LIME(Ribeiro et al., 2016). LIME is able to take any machine learning model, independent of its architecture, and explain the contribution of each feature on the final result. To understand the cause behind poor performance, particularly on visual data, LIME was used on the models developed in this work. The explanations in Figure 18 and 19 are obtained from the best performing



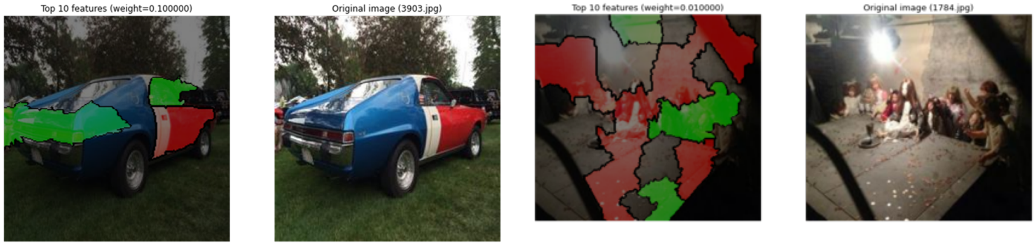
```
Id: 3653
RT @ZOO: .@RosieMacck took a shower, jaws-dropped. We were left speechless too!
Label: positive
```

took shower jawsdropped left speechless !

```
Id: 527
RT @ryansparkytay: Looks like someone got rejected... a moment of silence pls
Label: negative
```

looks like someone got rejected moment silence pls

Figure 18: LIME explanations for a positive and negative samples, respectively



Top 10 features (weight=0.100000)
Original image (3903.jpg)
Top 10 features (weight=0.010000)
Original image (1784.jpg)

Figure 19: LIME explanations for a positive and negative samples, respectively

models for each modality.
Looking at Figure 18, the model shows capability of understanding the sentiment of some words. In the second example, words like ”rejected” and ”silence” are correctly identified as negative. However, ”looks” and ”moment” are considered positive, despite not contributing towards that sentiment.
Figure 19 shows the explanations behind image predictions. Analysing these images, it is clear that the model does not show very good understanding of which features influence sentiment. In the first example, the car is what makes the image positive. However, this is not understood by the model as it only considers seemingly random parts of the car in its prediction. The same problem can be observed in the second picture where the walls influence the negative prediction despite being irrelevant.

5 Conclusions and future work

Since the invention of the internet, it is very simple for anyone to share their thoughts with the world. With this information, sentiment analysis can be a useful tool for understanding public opinion about a given topic.

The focus of this work was to implement various models with the capability to correctly identify sentiment in single and multimodal data.

To obtain the best performance from the models, many different hyperparameters were tested. Many natural language processing techniques were used for processing text. In the case of images, many convolutional neural network architectures and processing techniques present in the literature were tested.

Model tests were first performed on individual modalities. Afterwards, the best models from each modality were combined for analysing the multimodal case using an early fusion strategy.

Tackling the problem of bias, models were trained for each of the three annotators and results suggest that performance may be dependent on who annotates a particular dataset.

Then, the final results were obtained by testing with a majority vote strategy, and subsequently compared with the authors of MVSA. Both approaches showed a preference towards the positive class, although the models developed in this work displayed a stronger bias in the positive class, leading to more unbalanced results. Images in particular were harder to correctly classify. As shown by the LIME algorithm, this may be due to the image model's poor capability of understanding the most relevant features in an image.

There is still much work to be done in the field of sentiment analysis. Given the lack of data and the subjectivity of the problem, models have a hard time understanding sentiment. More research is needed, particularly in image and multimodal sentiment analysis.

the future, additional work may include the study of different techniques for better exploiting the content of images and the correlation between textual and visual features. Attention is a recent technique which allows models to focus on specific parts of an input. This may be particularly useful for visual sentiment analysis where certain features such as the background are often irrelevant and only add noise. The opinions present in text can influence the perceived sentiment in an image or video. Therefore, the link between modalities should be further explored. Identifying relevant features in each modality and analysing their relationship may improve multimodal performance.

References

U. Ahsan, M. De Choudhury, and I. Essa. Towards using visual attributes to infer image sentiment of social events. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 1372–1379, 2017. doi: 10.1109/IJCNN.2017.7966013.

Alexa, 2020. `https://www.alexa.com/topsites`.

Sabari Balaji, 2019. `https://medium.com/@cs.sabaribalaji/overfitting-6c1cd9af589`.

Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM international conference on Multimedia*, pages 223–232, 2013.

Kevin W. Bowyer, Nitesh V. Chawla, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: synthetic minority over-sampling technique. *CoRR*, abs/1106.1813, 2011. URL `http://arxiv.org/abs/1106.1813`.

Guoyong Cai and Binbin Xia. Convolutional neural networks for multimedia sentiment analysis. In *Natural Language Processing and Chinese Computing*, pages 159–167. Springer, 2015.

Meng Chen, Lin-Lin Zhang, Xiaohui Yu, and Yang Liu. Weighted co-training for cross-domain image sentiment classification. *Journal of Computer Science and Technology*, 32(4):714–725, 2017.

Minghai Chen, Sen Wang, Paul Pu Liang, Tadas Baltrusaitis, Amir Zadeh, and Louis-Philippe Morency. Multimodal sentiment analysis with word-level fusion and reinforcement learning. *CoRR*, abs/1802.00924, 2018. URL `http://arxiv.org/abs/1802.00924`.

Tao Chen, Damian Borth, Trevor Darrell, and Shih-Fu Chang. Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks, 2014a.

Tao Chen, Felix X. Yu, Jiawei Chen, Yin Cui, Yan-Ying Chen, and Shih-Fu Chang. Object-based visual sentiment concept analysis and application.

In *Proceedings of the 22nd ACM International Conference on Multimedia*, MM '14, page 367–376, New York, NY, USA, 2014b. Association for Computing Machinery. ISBN 9781450330633. doi: 10.1145/2647868.2654935. URL https://doi.org/10.1145/2647868.2654935.

Nadia FF Da Silva, Eduardo R Hruschka, and Estevam R Hruschka Jr. Tweet sentiment analysis with classifier ensembles. *Decision Support Systems*, 66: 170–179, 2014.

Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions?, 2016.

Datareportal, 2021. https://datareportal.com/social-media-users.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009.

Lingjia Deng and Janyce Wiebe. Mpqa 3.0: An entity/event-level sentiment corpus. pages 1323–1328, 01 2015. doi: 10.3115/v1/N15-1146.

Andrea Esuli and Fabrizio Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy, May 2006. European Language Resources Association (ELRA). URL http://www.lrec-conf.org/proceedings/lrec2006/pdf/384$_{p}$*df.pdf*.

António Gaspar and Luís Alexandre. A multimodal approach to image sentiment analysis. 2019. URL http://hdl.handle.net/10400.6/8162.

M. Ghiassi, J. Skinner, and D. Zimbra. Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with Applications*, 40(16):6266 – 6282, 2013. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2013.05.057. URL http://www.sciencedirect.com/science/article/pii/S0957417413003552.

Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, 1(12):2009, 2009.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL `http://arxiv.org/abs/1512.03385`.

I. Fogel I and D. Sagi. Biological cybernetics o springer-verlag 1989 gabor filters as texture discriminator.

Brendan Jou, Tao Chen, Nikolaos Pappas, Miriam Redi, Mercan Topkara, and Shih-Fu Chang. Visual affect around the world. *Proceedings of the 23rd ACM international conference on Multimedia*, Oct 2015. doi: 10.1145/2733373.2806246. URL `http://dx.doi.org/10.1145/2733373.2806246`.

Marie Katsurai and Shin'ichi Satoh. Image sentiment analysis using latent correlations among visual, textual, and sentiment views. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2837–2841, 2016. doi: 10.1109/ICASSP.2016.7472195.

Svetlana Kiritchenko, Xiaodan Zhu, and Saif Mohammad. Sentiment analysis of short informal text. *The Journal of Artificial Intelligence Research (JAIR)*, 50, 08 2014. doi: 10.1613/jair.4272.

Zuhe Li, Yangyu Fan, Weihua Liu, and Fengqin Wang. Image sentiment prediction based on textual descriptions with adjective noun pairs. *Multimedia Tools and Applications*, 77, 01 2018. doi: 10.1007/s11042-016-4310-5.

David G Lowe. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, volume 2, pages 1150–1157. Ieee, 1999.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

Avinash Navlani, 2019. `https://www.datacamp.com/community/tutorials/svm-classificatio`

S. Nemati and A. R. Naghsh-Nilchi. Exploiting evidential theory in the fusion of textual, audio, and visual modalities for affective music video retrieval.

In *2017 3rd International Conference on Pattern Recognition and Image Analysis (IPRIA)*, pages 222–228, 2017. doi: 10.1109/PRIA.2017.7983051.

Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El-Saddik. Sentiment analysis on multi-view social data. In *MultiMedia Modeling*, page 15–27, 2016.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL `http://www.aclweb.org/anthology/D14-1162`.

Anand Rajaraman and Jeffrey David Ullman. *Data Mining*, page 1–17. Cambridge University Press, 2011. doi: 10.1017/CBO9781139058452.002.

Joseph Redmon, Santosh Kumar Divvala, Ross B. Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. *CoRR*, abs/1506.02640, 2015. URL `http://arxiv.org/abs/1506.02640`.

Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. ”why should I trust you?”: Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016. URL `http://arxiv.org/abs/1602.04938`.

Y. Roh, G. Heo, and S. E. Whang. A survey on data collection for machine learning: A big data - ai integration perspective. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2019. doi: 10.1109/TKDE.2019.2946162.

Hassan Saif, Yulan He, Miriam Fernandez, and Harith Alani. Contextual semantics for sentiment analysis of twitter. *Information Processing & Management*, 52(1):5–19, 2016.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition, 2015.

Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15(84):2949–2980, 2014. URL `http://jmlr.org/papers/v15/srivastava14b.html`.

M. Sun, J. Yang, K. Wang, and H. Shen. Discovering affective regions in deep convolutional neural networks for visual sentiment prediction. In *2016 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2016. doi: 10.1109/ICME.2016.7552961.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. *CoRR*, abs/1512.00567, 2015. URL `http://arxiv.org/abs/1512.00567`.

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261, 2016. URL `http://arxiv.org/abs/1602.07261`.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. Learning sentiment-specific word embedding for Twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565, Baltimore, Maryland, June 2014. Association for Computational Linguistics. doi: 10.3115/v1/P14-1146. URL `https://www.aclweb.org/anthology/P14-1146`.

Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American society for information science and technology*, 61(12):2544–2558, 2010.

TOPJOBS, 2019. `https://www.topbots.com/solve-ai-nlp-problems-guide/bag-of-words/`.

Lucia Vadicamo, Fabio Carrara, Andrea Cimino, Stefano Cresci, Felice Dell'Orletta, Fabrizio Falchi, and Maurizio Tesconi. Cross-media learning for image sentiment analysis in the wild. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 308–317, Oct 2017. doi: 10.1109/ICCVW.2017.45.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

Wikipedia, 2014. https://en.wikipedia.org/wiki/Precision$_{a}nd_{r}ecall$.

Jie Xu, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. Visual-textual sentiment classification with bi-directional multi-level attention networks. *Knowledge-Based Systems*, 178:61–73, 2019a. ISSN 0950-7051. doi: https://doi.org/10.1016/j.knosys.2019.04.018. URL https://www.sciencedirect.com/science/article/pii/S0950705119301911.

Jie Xu, Feiran Huang, Xiaoming Zhang, Senzhang Wang, Chaozhuo Li, Zhoujun Li, and Yueying He. Sentiment analysis of social images via hierarchical deep fusion of content and links. *Applied Soft Computing*, 80:387–399, 2019b. ISSN 1568-4946. doi: https://doi.org/10.1016/j.asoc.2019.04.010. URL https://www.sciencedirect.com/science/article/pii/S1568494619302017.

Nan Xu, Wenji Mao, and Guandan Chen. *A Co-Memory Network for Multimodal Sentiment Analysis*, page 929–932. Association for Computing Machinery, New York, NY, USA, 2018. ISBN 9781450356572. URL https://doi.org/10.1145/3209978.3210093.

Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang. Robust image sentiment analysis using progressively trained and domain transferred deep networks, 2015.

Jianbo Yuan, Sean Mcdonough, Quanzeng You, and Jiebo Luo. Sentribute: image sentiment analysis from a mid-level perspective. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining*, pages 1–8, 2013.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. MOSI: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *CoRR*, abs/1606.06259, 2016. URL http://arxiv.org/abs/1606.06259.

Lei Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu. Combining lexicon-based and learning-based methods for twitter sentiment analysis. *HP Laboratories, Technical Report HPL-2011*, 89, 2011.