Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 1469–1477
Marseille, 11–16 May 2020

Aspect Flow Representation and Audio Inspired Analysis for Texts

Larissa L. Vasconcelos[1,2], Claudio E. C. Campelo[1], Caio L. M. Jeronimo[1]
Federal University of Campina Grande[1], Federal Institute of Paraiba[2]
882, Aprigio Veloso, Campina Grande - PB - Brazil[1]
PB-264, Monteiro - PB - Brazil[2]
larissa.vasconcelos@ifpb.edu.br, campelo@computacao.ufcg.edu.br, caiolibanio@copin.ufcg.edu.br

Abstract

For better understanding how people write texts, it is fundamental to examine how a particular linguistic aspect (e.g., subjectivity, sentiment, argumentation) is exploited in a text. Analysing such an aspect of a text as a whole (i.e., through a summarised single feature) can lead to significant information loss. In this paper, we propose a novel method of representing and analysing texts that consider how an aspect behaves throughout the text. We represent the texts by aspect flows for capturing all the aspect behaviour. Then, inspired by the resemblance between these flows format and a sound waveform, we fragment them into frames and calculate an adaptation of audio analysis features, named here Audio-Like Features, as a way of analysing the texts. The results of the conducted classification tasks reveal that our approach can surpass methods based on summarised features. We also show that a detailed examination of the Audio-Like Features can lead to a more profound knowledge about the represented texts.

Keywords: Text Representation, Text Analysis, Aspect Flows

1. Introduction

A challenge in Natural Language Processing (NLP) is to reproduce the way humans perceive a certain linguistic aspect in a text and, with that, achieve a better understanding of how different kinds of texts are written. For instance, analysing how fake news make use of subjective language can lead to meaningful knowledge about them (Jeronimo et al., 2019).

In this study, we employ the term *linguistic aspect* (or just *'aspect'*, for short) to refer to a distinguishable linguistic characteristic expressed in a text, such as sentiment, argumentation, or subjectivity. Observing the particularities of how a text utilises some aspect requires not only a global analysis of the entire text or individual analysis of each word or sentence, but also an analysis of its behaviour throughout the text.

Generally, for executing NLP tasks, the applied techniques model local representations of an aspect to extract summarised features that represent the input text as a whole, frequently by an average or median of that local representations. This sort of representation can lead to relevant information loss, especially for large texts, as they can be ignoring significant aspect singularities present in any part of the text that could be decisive on text type identification and characterisation (Aker et al., 2019).

An approach to avoiding representing a text in a globally summarised way is modeling a text as a flow, which is defined by Mao and Lebanon (2007) as a sequence of information collected from the words, sentences, or paragraphs of the text. In their study, Mao and Lebanon (2007) use sentiment flows to represent texts by assigning for each sentence one of the following values: 2 (highly praised), 1 (something good), 0 (objective description), -1 (something that needs improvement) and -2 (strong aversion). They propose a variant of conditional random fields (Lafferty et al., 2001) to proceed local and global sentiment prediction in reviews using the entire flows as features. Wachsmuth and Stein (2017) represent the text's discourse-level structure as a flow of rhetorical moves. They model until four kinds of text segment flow: local sentiment, modeling negative, neutral, and positive sentiment; discourse relation between segments, e.g., cause, circumstance, condition; paragraph-level discourse functions (introduction, body, rebuttal, conclusion); and argument roles, modeling real arguments, premises or claims. They propose a clusterisation in training flows and compare test flows to the training cluster's centroids in order to perform global reviews sentiment classification and essay scoring. Filatova (2017) models product reviews as sentiment flows and uses sentiment changing for sarcasm detection. She uses the Stanford Sentiment Analysis tool (Socher et al., 2013) with the 5-point sentiment scale (very negative (-2), negative (-1), neutral (0), positive (+1), very positive (+2)) to assign sentiment labels to each sentence in texts. Lee et al. (2010) represent texts as a merge of a sentiment flow and a relevance flow (defined in Seo and Jeon (2009)) to proceed with opinion retrieval. For each sentence of the text, they calculate a score that reflects its relevance (concerning a query) and opinion (the frequency of a lexicon's opinion words). As features, they use the variance of sentence scores, the fraction of peaks, and the first peak position.

In this paper, we propose to represent texts as aspect flows and perform a sophisticated flow analysis based on the concept of frame from audio analysis. The solution is independent of the target aspect being investigated, so its selection depends on what kind of information is meaningful to a given NLP task. In order to obtain an informative manner of analysing the way the aspect behaves throughout the text, our proposed approach divides the aspect flows into frames. Then, it extracts the so-called Audio-Like Features, an adaptation of audio analysis features for the text-domain. We evaluated the model in three NLP classification tasks: Fake News Classification Based on Text Subjectivity, Newspaper Columns Classification Based on Text Subjec-

tivity, and Movie Reviews Sentiment Classification. The first task uses subjectivity flows to explore the differences between legitimate and fake news, since fake news is more subjective than legitimate news (Jeronimo et al., 2019). In the second task, we also generate subjectivity flows, however, the aim is to differentiate objective news from newspaper columns, since newspaper columns are opinionated texts and objective news should not be. The latter task, on the other hand, investigates the overall movie reviews sentiment by analysing flows made of sentiment polarities.

Through the evaluation tasks, the proposed model reveals to be a viable form to represent and analyse texts, providing meaningful features for examining how a given aspect behaves throughout the texts, making it feasible to acquire valuable knowledge about the subject tasks.

The rest of the paper is structured as follows. Section 2. describes the proposed text representing and analysing model. Following, in Section 3., we explore in detail the executed experiments, including used datasets, lexicons, and experimental setup, as well as these experimental results and discussion. Finally, the paper concludes with Section 4., which depicts the conclusions drawn from the evaluation and outlines the possible future lines of work.

2. Proposed Model

This section introduces our proposed model, describing the aspect flows generation, the division of flows into frames, and the subsequent extraction of the features inspired by audio analysis, the audio-like features. Figure 1 shows a diagram of the proposed model.

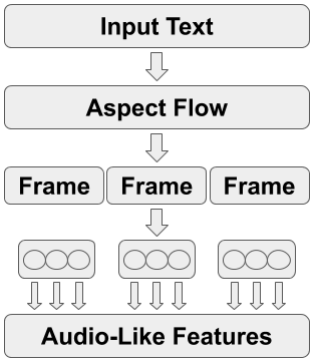

Figure 1: Model Diagram.

2.1. Aspect Flows Generation

Representing texts by a flow of an aspect related to a task is a promising way to better understand how the aspect behaves in the text, as the sentiment flow modeling explored in Mao and Lebanon (2007) article confirms.

For generating the aspect flow representation of the text, first it is necessary to split the text into sentences and then obtain an aspect representation for each one. The flow is the sequence of these sentence aspect representations. The aspect representation of a sentence can be generated, for instance, through a model trained on an annotated dataset, or via a model based on semantic similarity computation between the sentence and an aspect lexicon. All three tasks performed in this paper use the latter method to construct the flows, making annotated bases not necessary.

2.2. Audio-Like Features Extraction

If we plot an aspect flow as a graphic using the x-axis to represent the sentences and the y-axis the aspect values, we can perceive similarities between the form of this plot and a graphic of a sound waveform. In order to illustrate that, Figure 2 shows an example of an argumentation subjectivity flow of a legitimate news in our dataset. Given the

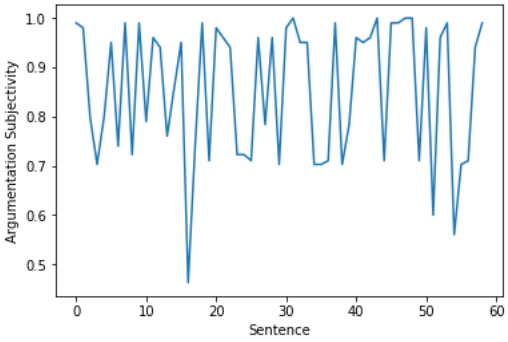

Figure 2: Example of an Argumentation Subjectivity Flow Graphic.

above, we propose to adapt the manner audio analysis is performed to the text's domain. Audio analysis is a solidified research area focused on providing useful knowledge about audio content (Giannakopoulos and Pikrakis, 2014a). This knowledge has proved valuable in many fields, such as segmentation and classification for music recommendation (Lu and Jiang, 2002; Helmholz et al., 2019), speech-music classification (Bhattacharjee et al., 2018), song emotion analysis (Jamdar et al., 2015), and sentiment extraction from speech streams (Kaushik et al., 2013).

2.2.1. Aspect Flow Frames

Frequently, in audio analysis, the audio signal is broken into possibly overlapping short-term windows (frames), and the analysis is carried out on a frame basis. In order to understand the reason for using this windowing technique, consider one audio which presents a conversation and a gunshot in the middle of that. If we compute an average intensity of the samples of the whole recording, the samples presenting the gunshot will dominate the result. If we analyse just this metric, we can obtain disturbed conclusions about the audio. Hence, it seems more feasible to compute features from the audio frames to better represent the information present there (Giannakopoulos and Pikrakis, 2014b).

As we aim to examine how aspects behave throughout texts, we propose to adopt the short-term windowing technique, fragmenting aspect flows into, initially, non-overlapping frames. In order to be able to compare the same parts of different texts (which very often have different sizes), our model breaks the aspect flows in a fixed number of frames. Therefore, regardless of the number of sentences in a text, the first frame will represent the first part of the text, for example, which we can compare to another text's first

part. Concerning defining the number of frames to split the flows, it is a dataset-dependent decision, as, for instance, if we are dealing with books, we can obtain so much more meaningful frames than with movie reviews.

2.2.2. Aspect Flow Audio-Like Features

In audio analysis, there are two categories of frame extracted features: time-domain and frequency-domain. The time-domain features offer a simple way to analyse audio signals and are directly extracted from the samples of the audio signal (waveform). On the other hand, frequency-domain features are extracted from the sound spectrum, a representation of the distribution of the frequency content of sounds (Giannakopoulos and Pikrakis, 2014d). Obtaining this representation requires to compute the Discrete Fourier Transform (DFT) of the audio signal (Giannakopoulos and Pikrakis, 2014c).

Our model presents the Audio-Like Features, an adaptation of audio analysis features extracted from the texts' aspect flows. Initially, it implements only the time-domain feature extraction, since it is possible to perform it directly from the flows. Our three Audio-Like Features are Energy, Median-Crossing Rate, and Energy Entropy, and will be detailed hereafter.

The first Audio-Like Feature is Energy. As the original version, this feature reflects the total magnitude of the aspect in the flow (Jalil et al., 2013). Let $x_i(n), n = 1, ..., F_L$ be the sequence of sentences of the i-th aspect frame, where F_L is the length of the frame. The implementation of Energy is defined as:

$$E(i) = \frac{1}{F_L} \sum_{n=1}^{F_L} |x_i(n)|^2 \quad (1)$$

Here we normalised the Energy by dividing it by F_L to remove the dependency on the frame length. The stronger an aspect appears in the frame, the bigger the frame's Energy.

Median-Crossing Rate (MCR) is the adaption of audio frame feature Zero-Crossing Rate (ZCR), which is the rate of sign-changes of the signal during the frame. As the audio signal waveform amplitude varies from -1 to 1, the ZCR is the number of times the signal changes value, from positive to negative and vice versa, divided by the length of the frame [livro cap4]. As we cannot expect all the sort of aspect inputs to be in the same audio signal amplitude's range $[-1, 1]$, our implementation uses the aspect flow median ($flowmedian$) as the parameter to calculate the crossing rate. The MCR is defined according to the following equation:

$$MCR(i) = \frac{1}{2F_L} \sum_{n=1}^{F_L} |msgn[x_i(n)] - msgn[x_i(n-1)]| \quad (2)$$

where $msgn$ is a modification of sign function, the Median Sign Function, denoted by:

$$msgn[x_i(n)] = \begin{cases} 1, & \text{if } x_i(n) > flowmedian. \\ -1, & \text{if } x_i(n) < flowmedian. \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

MCR can be interpreted as a measure of the target aspect noisiness of a flow; in other words, it reflects the level of aspect variation in a flow frame.

The last aspect feature is the Entropy of Energy, which can be interpreted as a measure of abrupt changes in the energy level of an aspect flow (like in audio analysis (Giannakopoulos and Pikrakis, 2014d)). For example, it can detect if a frame presents sentences with deeply different levels of subjectivity. In order to extract it, we first divide each flow frame into K sub-frames. This parameter must not be less than $K = 2$ to ensure there will be at least 2 sub-frames. Hence, it depends on the mean frame size of the aspect flows generated from a dataset, so that the model can avoid the unwanted effect of generating various sub-frames in a short frame. Then, for each sub-frame j, we compute its Energy as in (1) and divide it by the total frame Energy, $Eframe_i$. The division is necessary to treat the resulting sequence of sub-frame energy values, $e_j, j = 1, ..., K$, as a sequence of probabilities, as in (4):

$$e_j = \frac{Esubframe_j}{Eframe_i} \quad (4)$$

where

$$Eframe_i = \sum_{k=1}^{K} Esubframe_k \quad (5)$$

At a final step, the entropy, $Ent(i)$ of the sequence e_j is computed according to the equation:

$$Ent(i) = -\sum_{j=1}^{K} e_j * log_2(e_j) \quad (6)$$

The more significant changes the frame presents, the lower the Entropy Energy resulting value is.

As we can notice in Equations (2) and (6), to calculate MCR and Energy Entropy correctly, the flow must contain, at least, 2 sentences per frame. Thus, this minimum requirement must be considered during the definition of the number of frames and the K parameter (as it requires, at least, one sentence per subframe). Considering this, our model is not appropriate to analyse tiny texts, such as those microblogs.

3. Experimental Evaluation and Discussion

In this section, we describe our experimental evaluation conduction, also presenting the results of the three evaluation tasks and a discussion about these obtained results. For each task, we describe the utilised datasets, the linguistics resources used to generate the aspect flows, the evaluation protocol, the performance measures, and the classification models.

3.1. Fake News Classification Based on Text Subjectivity

The need for fake news detection is clear and present given the massive dissemination allowed by social media and messaging applications and its consequences[1].

Jeronimo et al. (2019) demonstrated good results in performing Fake News Classification Task in a dataset of Brazilian legitimate and fake news, considering that their

[1] https://www.dw.com/en/brazil-police-to-probe-allegations-of-electiondisinformation-on-whatsapp/a-45965369

subjectivity levels are significantly different. For that, the authors rely on a set of subjectivity lexicons built by Brazilian linguists (Amorim et al., 2018) and build subjectivity feature vectors for each news article. For generating these feature vectors, the Word Mover's Distance (WMD) (Huang et al., 2016) between each news sentences and these lexicons considering the embedding the news words lie in is calculated. Then, an average of the distances of each document sentences to each lexicon is computed. As Jeronimo et al. (2019) use a summarised way to represent text aspects (average of sentences' WMD), we decided to evaluate our model by replicating their most challenging experimental scenario, which consists of fake news classification regardless of the domain and sources of legitimate and fake news.

3.1.1. Dataset

In this paper, we use the same dataset as Jeronimo et al. (2019), but we added 26 to the initial 95 fake news. The dataset of legitimate news was collected from two of the biggest news sites in Brazil, that are Estadao[2] and Folha de Sao Paulo[3]. The dataset has a total of 207,914 legitimate news, from the years 2014 to 2017, divided into different domains: Politics, Sports, Economy, and Culture. The fake news dataset is composed of 121 fact-checked fake news that strongly disseminates in Brazil, from the years of 2010 to 2017. These news were collected from two popular fact-checking services, that are e-Farsas[4] and Boatos[5]. The fake news dataset is formed by a total of 121 fake news from more than 40 news sources.

3.1.2. Subjectivity Lexicons

As the source of aspect information, we employ the same five Brazilian Portuguese subjectivity lexicons (Amorim et al., 2018) Jeronimo et al. (2019) did. These lexicons were built by Brazilian linguists and are described next:

- The argumentation dimension represents words and expressions that are related to a more argumentative discourse. Such discourse is often used when someone is trying to convince another person of a specific point of view.
- The presupposition dimension encompasses terms that are related to a previous assumption of something. This kind of discourse is mainly used in situations where the interlocutor assumes something as true, even when this is not the case.
- The sentiment lexicon contains words and terms related to emotional discourse. Such terms are also used in the context of fake news when the writer of the article tries to emotionally engage the reader.
- The valuation dimension expresses words related to the amount or intensification of something.
- The modalization discourse is used when the interlocutor has an established stance about something or someone.

[2]https://www.estadao.com.br/
[3]https://www.folha.uol.com.br/
[4]http://www.e-farsas.com/
[5]http://www.boatos.org/

3.1.3. Experiment

The main objective of the experiment is to evaluate how effective are the subjectivity flow representation and Audio-Like Features analysis for fake news classification. For the purpose of building the subjectivity flows, we use the same method as Jeronimo et al. (2019): for each news, we calculate its sentences' WMD to the five subjectivity lexicons considering the embedding the news words lie in. The main difference is that, instead of using an average of these WMD values, we use the sequence of them as an aspect flow to represent the news. Therefore, for each news in the dataset, we generate five subjectivity flows, one to each lexicon. Then, we fragment each flow in frames and calculate the three Audio-Like Features for every frame of the flow. We use all legitimate news, regardless of the domain and sources, and the fact-checked fake news as model input data. This is a challenging scenario because the legitimate and fake news are a mix of different domains and sources.

3.1.4. Experimental Setup

We evaluate the average number of sentences per document in the dataset to define the number of frames the subjectivity flows should be split into, and the value that should be assigned to the K parameter to calculate the Energy Entropy. Legitimate and fake news contain an average of 21 and 14 sentences per document, respectively. Thus, we decided to split the flows into 3 frames, resulting into 7 and 4.67 sentences per frame, on average, for the legitimate and fake news, respectively. Since we have obtained frames with few sentences on average, we decided to use $K = 2$, to have at least two sentences per sub-frame, on average, the minimal necessary number to calculate the Energy Entropy correctly. Considering this decision, many documents did not meet the minimum requirement, generating missing values in various cells when the features were calculated. We then decided to input these cells with the average of the corresponding feature values to proceed with the experiment.

To evaluate the applicability of our proposed features for classifying fake news and to compare our results to those of Jeronimo et al. (2019), we used the Random Forest and XGBoost models. As the dataset of legitimate news is far more significant than the fake one, we randomize the train/test executions by varying the legitimate news documents 500 times. We also follow the proportion of four legitimate news to one fake news (Silverman, 2016). To calculate the semantic distances with WMD, we used the word embedding model from a large Wikipedia dump in Portuguese trained by Jeronimo et al. (2019). We evaluate the models in terms of the Area Under the Precision-Recall curve (PR-AUC), a metric that suits our scenario of class imbalance (Saito and Rehmsmeier, 2015; Davis and Goadrich, 2006).

3.1.5. Results and Discussion

We performed the classification task by training Random Forest and XGBoost classification algorithms with: (1) the Average Features proposed by Jeronimo et al. (2019), referred here as *'Average Model'*; (2) the Audio-Like Features of all three frames - *'All Frames Model'*; and (3) the Audio-Like Features per individual frame - *'Single Frame*

	AVG Features	AL Features All Frames	AL Features Frame 0	AL Features Frame 1	AL Features Frame 2
Random Forest PR-AUC	0.28 ± 0.03	0.60 ± 0.03	0.39 ± 0.04	0.53 ± 0.04	0.55 ± 0.03
XGBoost PR-AUC	0.27 ± 0.03	0.57 ± 0.04	0.37 ± 0.04	0.51 ± 0.05	0.53 ± 0.03

Table 1: Average PR-AUC results for the models trained with Jeronimo et al. Average Features (AVG Features), Audio-Like Features of all frames (AL Features All Frames) and Audio-Like Features per frame (AL Features Frame $\{0,1,2\}$).

Model'. The average of the PR-AUC for all 500 runs for each trained model is shown in Table 1. It is possible to visualize that the All Frames Model outperformed the Average Model by $\approx 53\%$ with Random Forest and $\approx 52.6\%$ with XGBoost. Moreover, it can be noticed that all Single Frame Models obtained better results than the Average Model, especially the Single Frame 2 Model that outperformed it by $\approx 49\%$ with both classification algorithms. This model also outperforms the others Single Frame Models, however it achieved almost the same result as the Single Frame 1 Model, indicating that the subjectivity aspect tends to be more decisive in differentiating legitimate and fake news in the middle and, specially, in the last portion of the texts. These results show that representing texts with aspect flows and analysing them using Audio-Like Features can improve the power of classification and possibly point the text's excerpts that presents more meaningful information referring to the task.

3.1.6. Audio-Like Features Analysis

We proceed with some analysis of the Audio-Like Features throughout the three frames, to exemplify what kind of information about the dataset our method can provide. Figure 3 shows the MCR boxplots of the presupposition flow for frames 0, 1 and 2. It can be seen that fake news values are higher than legitimate ones in all three frames. In addition, fake news MCR values increase throughout the frames, while legitimate news MCR values remain stable. These findings show that fake news is more unstable related to the presupposition lexicon distances than legitimate news, mainly in the last portion of the text. The boxplots of the Energy Entropy for the argumentation flow are shown in Figure 4. It can be observed that fake news present higher values than legitimate news in the first frame, and then this situation changes in the other frames. This information means that, in the beginning of the text, legitimate news undergo more abrupt changes in the WMD to the argumentation lexicon, whereas such abrupt changes occur in the middle and in end of the text in the case of fake news.

3.2. Newspaper Columns Classification Based on Text Subjectivity

Column is a recurring feature written by the same author in a newspaper. It is often characterised by the voice, personality, and opinions of the writer, in opposition to objective news that reports facts. Considering these characteristics, it seems feasible to discriminate between objective news and newspaper columns based on subjectivity of the language used in texts. Hence, we evaluate the task of classifying Newspaper Columns Classification Based on Text Subjectivity, by adopting a methodology similar to that presented in Section 3.1., based on the WMD to subjectivity lexicons.

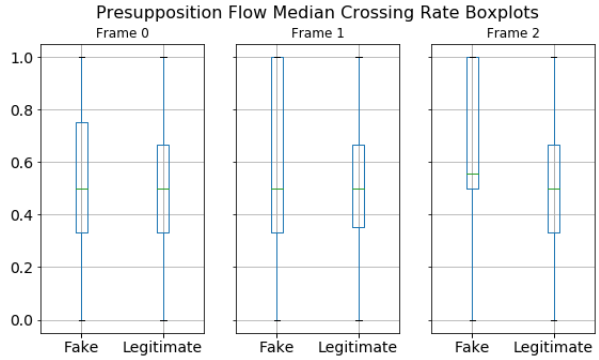

Figure 3: Boxplots of the Median Crossing Rate (MCR) for the Presupposition Flow.

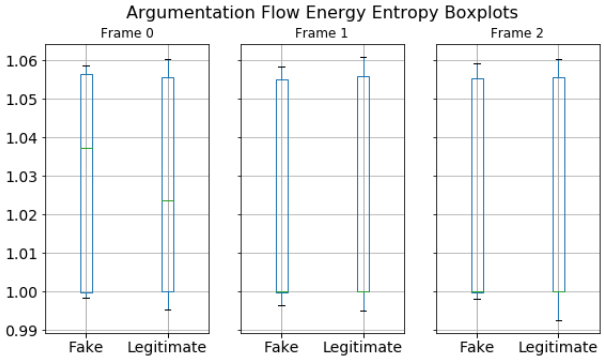

Figure 4: Boxplots of the Energy Entropy for the Argumentation Flow.

3.2.1. Dataset

We use the legitimate news dataset, presented in Section 3.1.1., to represent the objective news, as they are not collected from columns. The newspaper columns dataset was collected from Folha de Sao Paulo[6]. We collected a total of 7,062 columns articles, from a variety of domains (e.g., politics, economy, business, tourism) from 2010 to 2018.

3.2.2. Experiment

The main objective of this experiment is to evaluate how effective are the subjectivity flow representation and the Audio-Like Features analysis for objective news versus newspaper columns classification. The flows generation and Audio-Like Features calculation follow the same steps described in Section 3.1.3.. We use all the objective news and columns articles, regardless of the domain and sources, to keep the challenging character of the task.

[6]https://www.folha.uol.com.br/

3.2.3. Experimental Setup

As above mentioned, the objective news presents an average of 21 sentences per document, but the newspaper columns dataset presents an average of 29 sentences. Despite having a more significant number of sentences average, the newspaper columns dataset present lots of smaller texts, so we decided to maintain the number of 3 frames; therefore, we have 7 and 9.67 sentences per frame, on average, for the objective news and newspaper columns dataset, respectively. Additionally, we have maintained $K = 2$ to ensure there will be at least three sentences per sub-frame, in order to evaluate the model using bigger sub-frames to calculate the Energy Entropy. In spite of this, a lot of documents do not yet meet the minimal requirements, and therefore we had to input the features average. We also kept the other setup guidelines,such as the use of the Random Forest and XGBoost models for evaluating the applicability of our model compared to the Average Features. Although the newspaper columns dataset is more significant than the fake news dataset, it is still far less significant than the objective one, therefore we keep the 500 times randomisation and the four to one proportion. The Area Under Precision-Recall curve (PR-AUC) remains the metric used to evaluate the models.

3.2.4. Results and Discussion

The average PR-AUC for all 500 runs of each trained model is shown in Table 2. Once more, the All Frames Model outperformed the Average Model (now the difference was of $\approx 34.6\%$ with Random Forest and $\approx 38.4\%$ with XGBoost). All Single Frame Models also performed better than the Average Model. The Single Frame 2 Model has surpassed the Average Model by $\approx 33\%$ with Random Forest and by $\approx 37\%$ with XGBoost. In this experiment, the Single Frame 2 Model achieved the best performance among all the Single Frame Models, which indicates that the subjectivity aspect is more efficient in discerning objective news from newspaper column in the ending excerpt of the texts. These results show that our proposed method can potentially improve the classification achievements, also pointing the most discriminative excerpt of the texts.

3.2.5. Audio-Like Features Analysis

Figure 5 shows the MCR boxplots of the sentiment flow for the 3 frames. Newspaper columns present smaller MCR values than objective news over all frames. Objective news maintains almost the same values throughout the frames, while column news presents a peak in the second frame. In other words, regarding the sentiment lexicon, the objective news is more unstable throughout the text, and newspaper columns present more instability in the middle of the text. Newspaper columns show smaller energy entropy values than objective news regarding the modalization flow in all frames, as the boxplots presented in Figure 6 confirm. From these analysis, we can conclude that newspaper columns undergo more abrupt changes in the WMD to the referred lexicon.

3.3. Movie Reviews Sentiment Classification

Opinionated information is widely available online and plays a vital role in evaluating whether a product or service is pleasing their consumers or not. In this context, sentiment analysis of product or service reviews is a common exploited field since it focuses on the classification of sentiments or opinions expressed in human-generated texts (Araque et al., 2019). In order to evaluate our proposed method performance against a summarised approach in several domains of texts and in different languages, we also conduct an experiment on sentiment classification of movie reviews written in English.

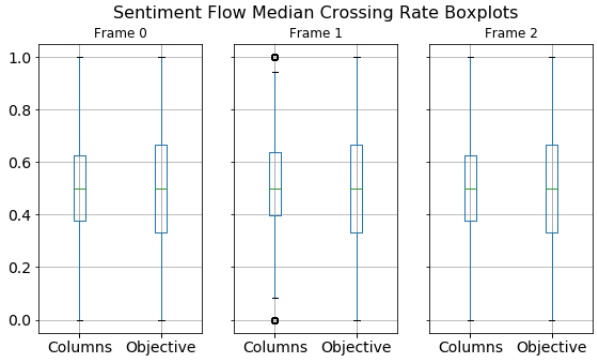

Figure 5: Boxplots of the Median Crossing Rate (MCR) for the Sentiment Flow.

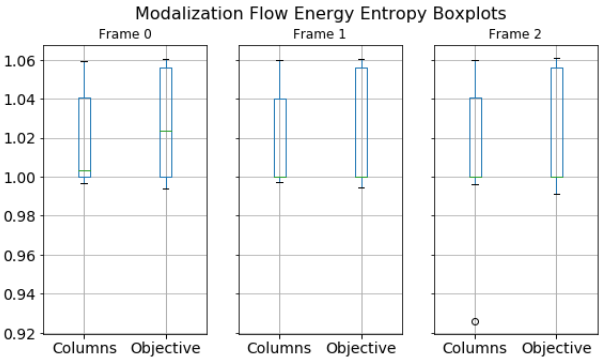

Figure 6: Boxplots of the Energy Entropy for the Modalization Flow.

3.3.1. Dataset

The dataset used in this task is the PL04 (Pang and Lee, 2004), containing 2,000 movie reviews written in English. There are 1,000 positive and 1,000 negative labeled movie reviews extracted from the IMDb[7] site of movies and TV shows reviews.

3.3.2. Sentiment Lexicon

With the purpose of constructing the sentiment flows, we calculate the WMD of the sentences of the PL04 texts to the sentiment lexicon AFINN (Nielsen, 2011). We used the newest version of AFINN, that contains a total of 2,477 English words, being 878 positives, and 1,578 negatives. Positive words are scored from 1 to 5, while negative ones have a sentiment score ranging from -5 to -1.

[7] https://www.imdb.com/

	AVG Features	AL Features All Frames	AL Features Frame 0	AL Features Frame 1	AL Features Frame 2
Random Forest PR-AUC	0.34 ± 0.01	0.52 ± 0.00	0.42 ± 0.00	0.45 ± 0.00	0.51 ± 0.00
XGBoost PR-AUC	0.32 ± 0.01	0.52 ± 0.00	0.42 ± 0.00	0.46 ± 0.00	0.51 ± 0.00

Table 2: Average PR-AUC results for the models trained with Jeronimo et al. Average Features (AVG Features), Audio-Like Features of all frames (AL Features All Frames) and Audio-Like Features per frame (AL Features Frame $\{0,1,2\}$).

	AVG Features	AL Features All Frames	AL Features Frame 0	AL Features Frame 1	AL Features Frame 2
F1 score	0.70 ± 0.03	0.72 ± 0.01	0.62 ± 0.04	0.64 ± 0.05	0.69 ± 0.03

Table 3: Average F1 score results for the models trained with Jeronimo et al. Average Features (AVG Features), Audio-Like Features of all frames (AL Features All Frames) and Audio-Like Features per frame (AL Features Frame $\{0,1,2\}$).

3.3.3. Experiment

In this case, we want to evaluate how beneficial are sentiment flow representation and Audio-Like Feature analysis for movie reviews sentiment classification. First, we separate the AFINN's negative from positive words, generating two polarity lexicons. Then we generate two sentiment flows (negative and positive) for each review, calculating the WMD to the lexicons. Afterwards, we proceed with the flow fragmentation into frames and the Audio-Like Features calculation, as performed in the other experiments.

3.3.4. Experimental Setup

We use the Logistic Regression model for evaluating the applicability of our approach, as this model shows quite effective results in executing this task (Araque et al., 2017). Concerning the training and test procedures, we have followed the PL04 associated cross-validation splits, which is composed of 10 splits, with 100 positive and 100 negative reviews each. We performed 10 executions in total, using 9 different splits to train and the remainder split to test the model in each execution. To calculate the semantic distances with WMD, we used the widely widespread pre-trained word vectors of Word2Vec approach[8]. To evaluate the models, we use the F1 score, a metric that seeks a balance between Precision and Recall (Araque et al., 2017). The length of reviews is of 30 sentences on average. However, this dataset contains several smaller texts. For this reason, we kept the division on 3 frames and defined $K = 2$. Considering these decisions, all the reviews in the dataset achieved the minimum requirement.

3.3.5. Results and Discussion

The average of the F1 score results of all the 10-fold runs of each trained model is shown in Table 3. We can figure out that the All Frames model obtains a better result than the Average Model, but not so expressively as in other scenarios. None of the Single Frame Models present a better F1 score than the Average Model. However, the Single Frame 2 Model, which represents the ending excerpt of the texts, is more successful in differentiating positive from negative reviews than the others. Analysing these findings, we can conclude that, even in this scenario that our approach does not achieve significantly better results than a summarised approach, it can suggest what portion of the text is more representative to a task.

3.3.6. Audio-Like Features Analysis

From the MCR boxplots of the positive flows shown in Figure 7, we can perceive that the negative reviews present higher MCR values than positive reviews in the first and second frames. In the third frame, the values are almost equal. The negative reviews instability regarding positive lexicon decreases from the first to the last frame, while the positive reviews show a peak in the second frame. If we consider the MCR boxplots of the negative flows (Figure 8), we realise that the situation is the opposite: the positive reviews instability is higher than the negative reviews and decreases over the frames.

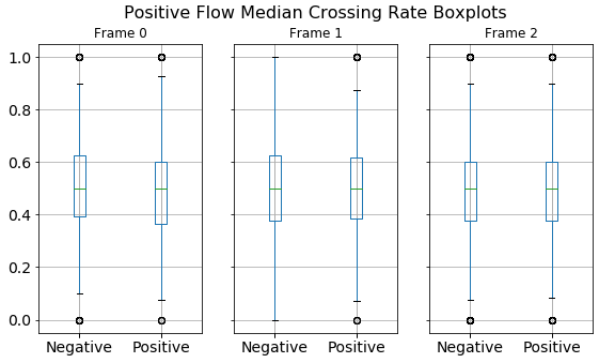

Figure 7: Boxplots of the Median Crossing Rate (MCR) for the Positive Flow.

4. Conclusions and Future Work

In this paper, we have introduced a promising novel approach to interpret human-generated texts representing and analysing them in their entirety, though not in a summarised way. More precisely, in order to represent texts, our model uses a sequence of information collected from the sentences of the text, which we called aspect flows. Then, inspired by audio analysis, this proposed model fragments the texts' aspect flows into frames and calculates Audio-Like Features for each one to perform text analysis. In the presented evaluation tasks, we have used aspect flows comprising the texts' sentences semantic distances to lexicons, considering

[8] https://code.google.com/archive/p/word2vec/

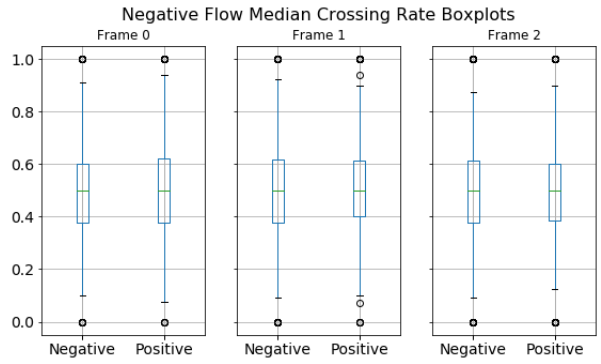

Figure 8: Boxplots of the Median Crossing Rate (MCR) for the Negative Flow.

a word embedding space. Using state-of-the-art machine learning classifiers, we have shown that this new approach outperforms the summarised features approach in different tasks that included diverse text kinds and domains and two distinct languages. Even when the results are not substantially better, our approach can evidence what portion of texts is more prone to differentiate them. Furthermore, we have shown that the investigation of the Audio-Like Features can reveal meaningful information about how each kind of text exploits an aspect, leading us to a deeper understanding of how these texts are written. As future work, we intend to implement frequency-domain features, after a criterion study about its viability, and mid-term features. We also plan to apply this method to other NLP tasks using larger texts.

5. Bibliographical References

Aker, A., Gravenkamp, H., Mayer, S., Hamacher, M., Smets, A., Nti, A., Erdmann, J., Serong, J., Welpinghus, A., and Marchi, F. (2019). Corpus of news articles annotated with article level subjectivity. 06.

Amorim, E., Cançado, M., and Veloso, A. (2018). Automated essay scoring in the presence of biased ratings. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 229–237, New Orleans, Louisiana, June. Association for Computational Linguistics.

Araque, O., Corcuera-Platas, I., Sanchez-Rada, J. F., and Iglesias, C. A. (2017). Enhancing deep learning sentiment analysis with ensemble techniques in social applications. *Expert Systems with Applications*, 77:236 – 246.

Araque, O., Zhu, G., and Iglesias, C. A. (2019). A semantic similarity-based perspective of affect lexicons for sentiment analysis. *Knowledge-Based Systems*, 165:346 – 359.

Bhattacharjee, M., Prasanna, S. R. M., and Guha, P. (2018). Time-frequency audio features for speech-music classification. *ArXiv*, abs/1811.01222.

Davis, J. and Goadrich, M. (2006). The relationship between precision-recall and roc curves. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 233–240, New York, NY, USA. ACM.

Filatova, E. (2017). Sarcasm detection using sentiment flow shifts. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017*, pages 264–269.

Giannakopoulos, T. and Pikrakis, A. (2014a). Chapter 1 - introduction. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 3 – 8. Academic Press, Oxford.

Giannakopoulos, T. and Pikrakis, A. (2014b). Chapter 2 - getting familiar with audio signals. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 9 – 31. Academic Press, Oxford.

Giannakopoulos, T. and Pikrakis, A. (2014c). Chapter 3 - signal transforms and filtering essentials. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 33 – 57. Academic Press, Oxford.

Giannakopoulos, T. and Pikrakis, A. (2014d). Chapter 4 - audio features. In Theodoros Giannakopoulos et al., editors, *Introduction to Audio Analysis*, pages 59 – 103. Academic Press, Oxford.

Helmholz, P., Meyer, M., and Robra-Bissantz, S. (2019). Feel the moosic: Emotion-based music selection and recommendation. 06.

Huang, G., Quo, C., Kusner, M. J., Sun, Y., Weinberger, K. Q., and Sha, F. (2016). Supervised word mover's distance. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pages 4869–4877, USA. Curran Associates Inc.

Jalil, M., Butt, F., and Malik, A. (2013). Short-time energy, magnitude, zero crossing rate and autocorrelation measurement for discriminating voiced and unvoiced segments of speech signals. pages 208–212, 05.

Jamdar, A., Abraham, J., Khanna, K., and Dubey, R. (2015). Emotion analysis of songs based on lyrical and audio features. *International Journal of Artificial Intelligence & Applications*, 6, 06.

Jeronimo, C., Marinho, L., Campelo, C., Veloso, A., and Melo, A. (2019). Fake news classification based on subjective language. In *Proceedings of the 21st International Conference on Information Integration and Web-based Applications & Services*.

Kaushik, L., Sangwan, A., and Hansen, J. (2013). Sentiment extraction from natural audio streams. 05.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, ICML '01, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Lee, S.-W., Lee, J.-T., Song, Y.-I., and Rim, H.-C. (2010). High precision opinion retrieval using sentiment-relevance flows. pages 817–818, 01.

Lu, L. and Jiang, H. (2002). Content analysis for audio classification and segmentation. *Speech and Audio Processing, IEEE Transactions on*, 10:504 – 516, 11.

Mao, Y. and Lebanon, G. (2007). Isotonic conditional random fields and local sentiment flow. In B. Schölkopf,

et al., editors, *Advances in Neural Information Processing Systems 19*, pages 961–968. MIT Press.

Nielsen, F. A. (2011). A new ANEW: evaluation of a word list for sentiment analysis in microblogs. *CoRR*, abs/1103.2903.

Pang, B. and Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*.

Saito, T. and Rehmsmeier, M. (2015). The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. In *PloS one*.

Seo, J. and Jeon, J. (2009). High precision retrieval using relevance-flow graph. In *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 694–695, New York, NY, USA. ACM.

Silverman, C. (2016). Hyperpartisan facebook pages are publishing false and misleading information at an alarming rate. buzzfeed, nov. 16.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA, October. Association for Computational Linguistics.

Wachsmuth, H. and Stein, B. (2017). A universal model for discourse-level argumentation analysis. *ACM Trans. Internet Technol.*, 17(3):28:1–28:24, June.