Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт (государственный университет)»
Физтех-школа прикладной математики и информатики
кафедра интеллектуальных систем

Направление подготовки: 03.03.01 Прикладные математика и физика (бакалавриат)
Направленность (профиль) подготовки: Компьютерные технологии и интеллектуальный анализ данных

Построение тематических моделей полилогов

(Бакалаврский диплом)

Студент:
Саттаров Тагир Ильдарович

(подпись студента)

Научный руководитель:
Воронцов Константин Вячеславович,
д. ф.-м. н.

(подпись научного руководителя)

Москва 2021

Оглавление

Аннотация

В данной работе анализируются полилоги теледебатов французских новостных каналов за период 2008-2020 гг с целью выявления признаков деградации качества дискуссии. В качестве основного признака выбрано среднее число слов в репликах участников дискуссии в предположении, что сокращение средней длины реплик свидетельствует о снижении качества дискуссии. На основе визуального анализа временных рядов данного признака делаются некоторые содержательные выводы о политике нескольких популярных телепередач в разные периоды времени. В качестве второго подхода выбран анализ тональности на основе готовых инструментов библиотеки Textblob, однако в экспериментах он не дал значимых результатов.

Ключевые слова: сентимент анализ, TextBlob, теледебаты, анализ полилогов

Глава 1

Введение

1.1 Введение

Отношение новостных каналов к текущим событиям порождает все больше и больше дискуссий. В настоящее время принято думать, что в гонке за аудиторию приоритет отдается не качеству и достоверности информации, а формату её подачи, и в связи с этим встречается выраженное использование конфронтаций. Но верно ли это распространённое убеждение? На данный момент не изучены количественные переменные, которые могли бы подтвердить или, по крайней мере, поддержать или опровергнуть эту мысль. Поэтому представляется целесообразным детально изучить содержание транслируемых по различным каналам передач, чтобы подтвердить или опровергнуть данную практику в средствах массовой информации при помощи подобранных маркеров[3].

Первым примером является новостной канал CNEWS, известный до 2017 года под названием Itélé и чей ребрендинг сопровождался приглашением скандальных персон[1][7]. Полемические выдержки из их шоу стали обычными, и канал бойкотировали многие бренды и личности. Тем не менее, аудитория CNEWS растет, в то время как данный сектор кажется насыщенным. Появление новых каналов таких как LCI и France Info на бесплатном телевещании в 2016-м году довело количество новостных каналов до четырех, и общественность начала все больше отходить от традиционных СМИ для получения информации.

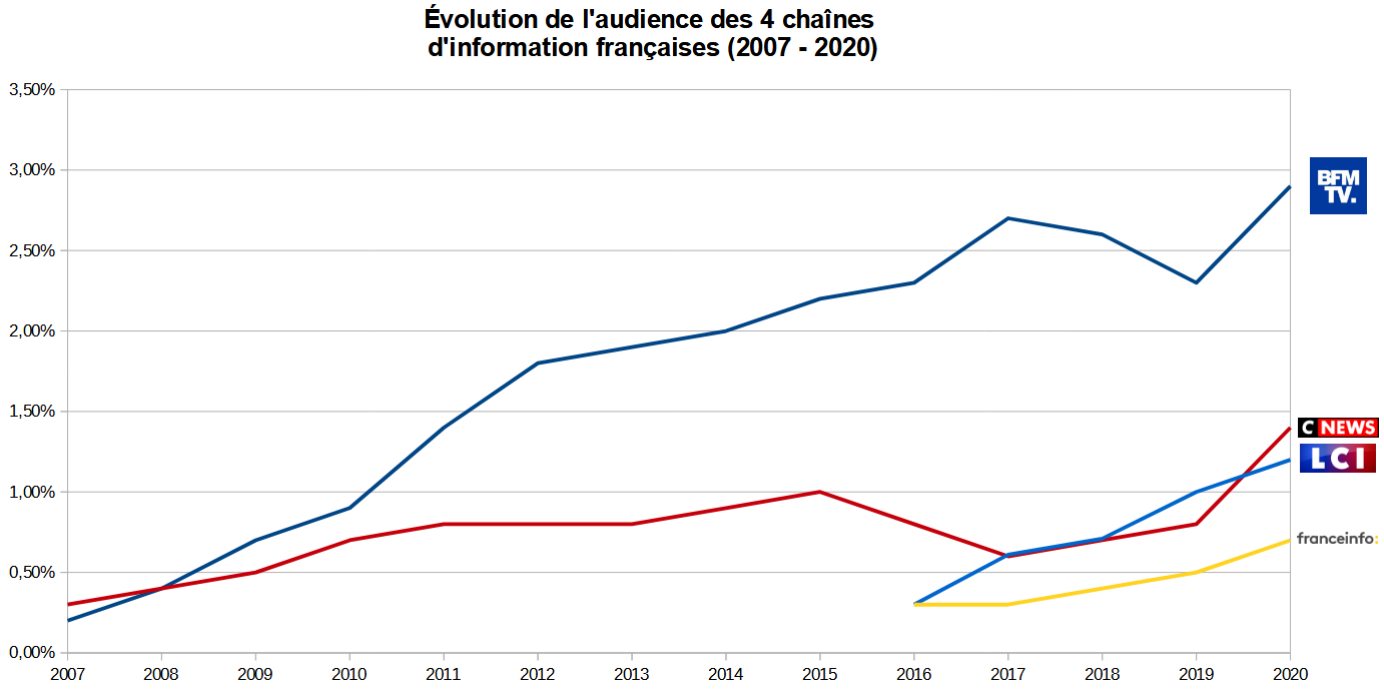

Рис. 1.1: Аудитория четырёх основных французских телеканалов. Источник Wikipedia

Итак, как мы можем количественно оценить изменения в стратегии телеканалов? К каким стратегиям стремились прийти информационные программы? Ответы на эти вопросы будут представлены в данной работе. Необходимость поиска индикаторов обусловлена огромным количеством данных и субъективностью его анализа отдельным человеком. Поэтому в данной работе производится поиск и верификация индикаторов, способных дать информацию об эволюции стратегии того или иного телеканала.

Для того чтобы найти ответ на поставленные вопросы, необходимо провести тщательную работу как с точки зрения выбора методов, которые должны быть реализованы, так и с точки зрения толкования результатов. Данное исследование страдает от некоторой предвзятости: мы начинаем с предположения, что качество дискуссии стало постепенно деградировать с течением времени, и что канал CNEWS является одним из ответственных за такое положение дел. Поэтому разрабатываются показатели, которые будут считаться актуальными, если они различают программы CNEWS и другие. Однако, выбор, который мы делаем, будет систематически обсуждаться, с тем, чтобы определить, влияют ли они на наше суждение по тем или иным программам.

Таким образом, в данном исследовании поставлены следующие задачи:

1. Определить источники данных и привести их структурную характеристику.

2. Провести анализ метаданных запрошенных передач и выявить границы применения

метода.

3. Выработать индикатор, основанный на анализе тональности и также выявить границы его применения.

Данная область на данный момент изучена недостаточно. Среди существующих исследований можно выделить работу по анализу транскрипций с нагрудных камер полицейских в США[4]. В работе исследуется отношение полицейских к чернокожим и белокожим гражданским. Здесь использован метод разметки данных добровольцами для того, чтобы дать количественную оценку вежливости полицейских по отношению к разным категориям граждан. Заметна схожесть с одной из поставленных целей нашего исследования– интересно отношение субъекта(в данном случае полицейского) ко второму участнику разговора. Также можно выделить методы анализа теледебатов[6], которые могут быть применены автоматически.

Глава 2

Данные

2.1 Данные

Источники данных Главной проблемой, возникающей в такого рода исследованиях, является доступ к достаточному объёму данных в удобном для изучения формате. Создание таких данных трудно представить с обычным доступом к ресурсам. Кроме того, обрабатываемый объём делает невозможным ручную оцифровку передач. Благодаря Этьену Олиону(Étienne Ollion), который помог нам заключить соглашение с l'Institut National de l'Audiovisuel (прим. Национальный аудиовизуальный институт, далее ИНА), был получен доступ к данным, удовлетворяющим критериям упомянутым во введении. Таким образом, ИНА предоставил свои архивы и сделал доступными выбранные программы. В общей сложности было расшифровано и передано более 14 000 передач, частично соответствующих сформулированным запросам.

Выпуск	**Кол-во**	**Канал**	**Суммарно**
Journal de 20 heures	4368	TF1	4368
On n'est pas couché	433	France 2	433
C à vous C dans l'air	2378 3353	France 5	5731
Ça se dispute Face à l'info L'heure des pros On ne va pas se mentir	544 268 1339 551	Itélé/CNEWS	2702
Le Débat	1310	LCI	1310
Les Grandes Gueules	522	RMC	522

Таблица 2.1: Распределение доступных телепередач по каналам.

Структурная характеристика данных До 2009 года доступные передачи практически отсутствуют. Поэтому наше исследование не будет включать период до 2009 года.

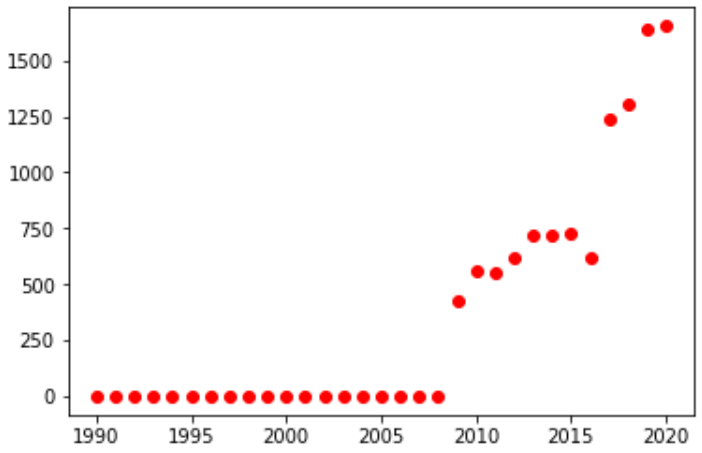

Рис. 2.1: Количество доступных выпусков по годам

Таким образом, остаётся возможным , в лучшем случае, получить представление об эволюции телепередач в средствах массовой информации в течение одиннадцатилетнего периода между 2009 и 2020 годами. С другой стороны, мы не сможем сравнить информационные каналы друг с другом, так как канал France Info основан относительно недавно (в 2016 году) и отсутствует в корпусе. Поэтому в рамках данного исследования было решено попытаться охарактеризовать:

- различия между CNEWS и LCI
- различия, существующие между программами одного и того же канала, в частности Itélé / CNEWS
- различия, которые существуют между передачами государственных каналов (France 2 и France 5) и передачами новостных каналов непрерывного вещания (CNEWS и LCI)
- различия, существующие между программами различных каналов (On n'est pas couché и Face à l'info) с участием одного и того же обозревателя Эрика Земмура

Случайный шум в данных Данные предоставлены INA в файлах формата CSV. Каждый файл имеет ряд речевых сегментов, каждый из них кодирует несколько слов. Используется простой алгоритм парсинга для извлечения сегментов и отдельных слов из сегментов. Первое ограничение при чтении файлов заключалось в том, что транскрипции передач созданы автоматически и, следовательно, содержат заметное количество ошибок, что делает некоторые предложения полностью бессмысленными. Кроме того, они воспроизводят ошибки и хезитации в речи, в том числе такие звуки как "euh"(хм, ээ. . .). В данной

работе предполагаем, что в среднем эти эффекты слабо сказываются на результатах. С другой стороны, эти ошибки могут быть более раздражающими при чтении и разметке данных непосредственно людьми.

```
<SpeechSegment ch="1" sconf="1.00" stime="830.820" etime="832.670"
spkid="S339" lang="fre" lconf="1.00" trs="1">
    <Word id="2871" stime="830.82" dur="0.17" conf="0.70"> que </Word>
    <Word id="2872" stime="831.09" dur="0.10" conf="0.41"> et </Word>
    <Word id="2873" stime="831.20" dur="0.28" conf="0.41"> eric </Word>
    <Word id="2874" stime="831.49" dur="0.35" conf="0.68"> menant </Word>
    <Word id="2875" stime="831.87" dur="0.48" conf="0.81"> raison </Word>
</SpeechSegment>
```

Рис. 2.2: Пример хаотичной транскрипции *« que et eric menant raison »*.

Ещё один факт, с которым мы имеем дело – время вещания телепередач и наличие рекламы. Здесь мы снова делаем предположение, что реклама длится незначительную часть передачи и влияет на результаты сравнений одинаково. Последнее может быть спорным, так как, например, государственные телеканалы не содержат рекламы по вечерам[8]. Также можно предположить, что тональность рекламы, как правило, позитивная, имеет тенденцию сдвигать общую тональность передач частных каналов, делая её более позитивной.

Глава 3

Вычислительный эксперимент

Как упоминалось выше, мы стремимся установить показатели, которые смогли бы помочь сравнивать программы и были устойчивы к шуму, присутствующему в данных. Также важна возможность оценивать влияние иных факторов на результаты(например, рекламы). Было определено два основных показателя, которые могут быть эффективно реализованы:

- анализ продолжительности предложений
- анализ тональности, цель которого заключается в установлении тональности выпусков

3.1 Использование метаданных

Метод

Для каждого выпуска создается соответствующий объект, который включает в себя, в дополнение к названию, дату, overall_index, соответствующий количеству дней, прошедших с 1 января 2008 года. Данная индексация позволяет сравнить выпуски друг с другом в хронологическом порядке. Объект также содержит поле, которое хранит среднее количество слов по сегментам выпуска. Затем мы при помощи библиотеки pandas Python[10][13] получаем следующую таблицу:

Название	**Дата**	**Кол-во слов**
20_heures	366	27.99570815450644
...	...	...
c_dans_l_air	4732	34.4676724137931

Таблица 3.1: Структура *DataFrame*.

Прежде чем интерпретировать полученные результаты, необходимо убедиться, что

сегменты соответствуют последовательным ответам и что транскриптор не упускает перебивания, которые мы хотели бы определить. Для этого мы рассмотрим программу L'heure des pros от 15 декабря 2020 года [13]. Далее следуют 6 сегментов, в которых каждый цвет, из трёх имеющихся, соответствует одному из участников передачи:

Segment 1. *je suis assez je suis assez séduits par votre théorie mais je pense qu' elle est elle va je pense que je vais vous savez c' est celle là qui qui est en marche citer le cas il faut le gouvernement dit aujourd'hui allez je vous demande un ultime effort c' est ce qui va se passer mais non respect de sauver avant quinze mais ça passera en trois semaines peut être disques mais*

Segment 2. *c' est faux espoirs mais tout passe huit mois après mois sont absolument dévastateur par ailleurs là où je vous rejoins c' est que notre état on s' aperçoit ne sert plus qu' à une chose c' est à distribuer des subventions c' est à dire que*

Segment 3. *quand on entend des discours il y a trois choses qui frappe*

Segment 4. *tout ce qui a été fait par l' état a échoué et maintenant on dit soyez raisonnable auto confinez vous papy et mamie à la cuisine fait les efforts qu' on est plus capable de faire pour vous protéger moi je voudrais pas demain*

Segment 5. *tout n' a pas souhaité que la l' argent qui coule à flots n' a pas échouer sauf que cet argent c' est que lui donne mais c' est le symbole de la faiblesse de l' état l' état de je vu mais tant mieux qu' il entend arrêté et en échange de sa l' état dépense et pour montrer qu' ils existent et c' est ce qui va se mais quand les restaurateurs échec mais je vous assure le lac saint c' est dans cette*

Segment 6. *je veux dire que je j' avais mis cette hypothèse je m' en souviens sur ce plateau ça veut dire que le le gouvernement est en train d' établir un vaccin obligatoire ça revient petit pas ça c' est à dire qu' il y a une grande hypocrisie à dire si vous ne faites vous faire pas vacciner vous ne sortirait pas*

Можно сделать вывод, что:

- реплика одного человека может быть разделена на несколько сегментов
- один и тот же сегмент может включать слова двух различных ораторов

В частности, перекрывающиеся реплики (например, в случае словесной перепалки) не обязательно будут обнаружены как сказанные различными лицами и будут сгруппированы в одном сегменте. Складывается впечатление, что алгоритм транскрипций уделяет

много внимания паузам при распознавании отдельных сегментов. Это объясняет разбиение длинных высказываний на несколько сегментов, при том, что алгоритм сохраняет информацию о том, что спикер не изменился. Далее, делается предположение, что все выпуски подвержены данному явлению в одинаковой степени и что случай Segment 5(рисунок выше) находится в меньшинстве. Тем не менее, короткие сегменты часто обусловлены лишь перебиванием. Длительные же вмешательства часто будут разделены между несколькими сегментами, но каждый сегмент будет значительной длины и будет являться полноценным предложением.

Предварительные выводы

Визуализация результатов для всех 14408 передач выглядит следующим образом:

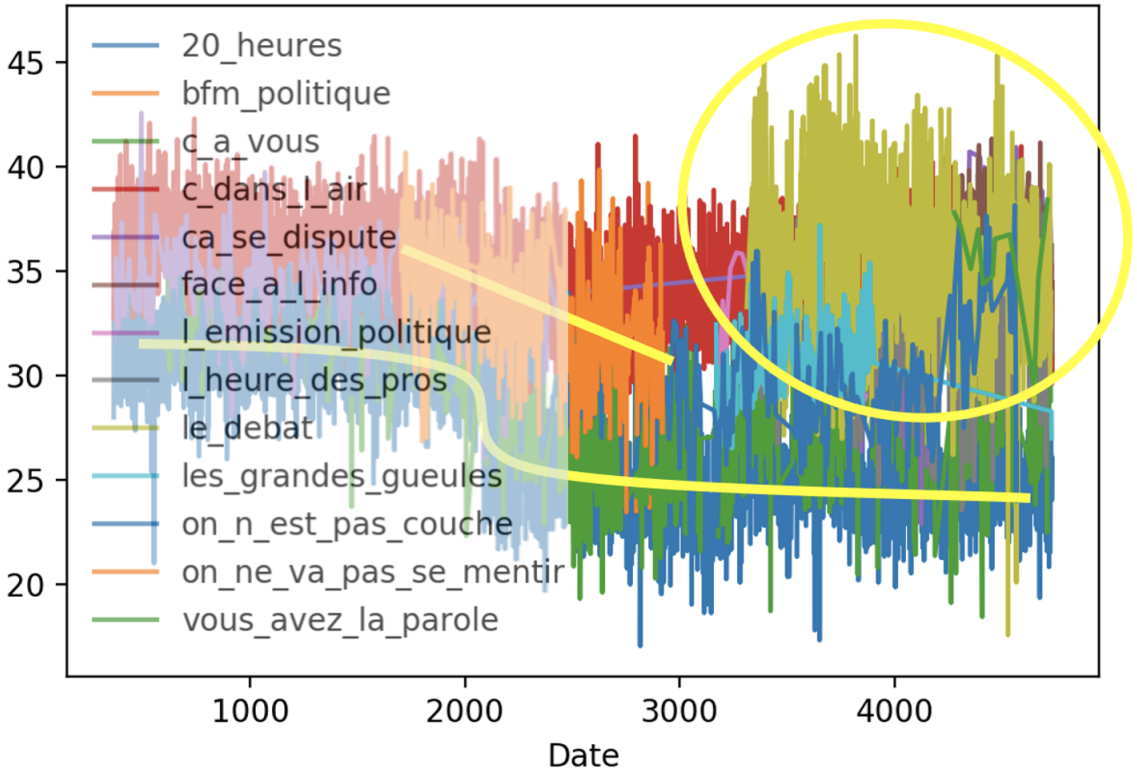

Рис. 3.1: Зависимость средней длины фразы различных телепередач от времени(кол-во дней с 01.01.2008)

Можно сделать несколько предварительных выводов:

1. программа C dans l'air (выделена красным цветом) выглядит без изменений с течением времени, в отличие от выпуска новостей TF1 (выделена синим цветом)

2. программа, рассмотренная для LCI, а именно Le Débat (в светло-зеленом цвете),

отличается более высоким средним показателем длины фразы

3. наблюдается тенденция к снижению средней длины фразы для передачи Le Débat. Это же характерно для передачи On va pas se mentir (оранжевый)

Таким образом, для некоторых передач прослеживается изменение данной характеристики во времени. Далее рассмотрим детальный анализ каждой из представленных передач.

Детальный анализ

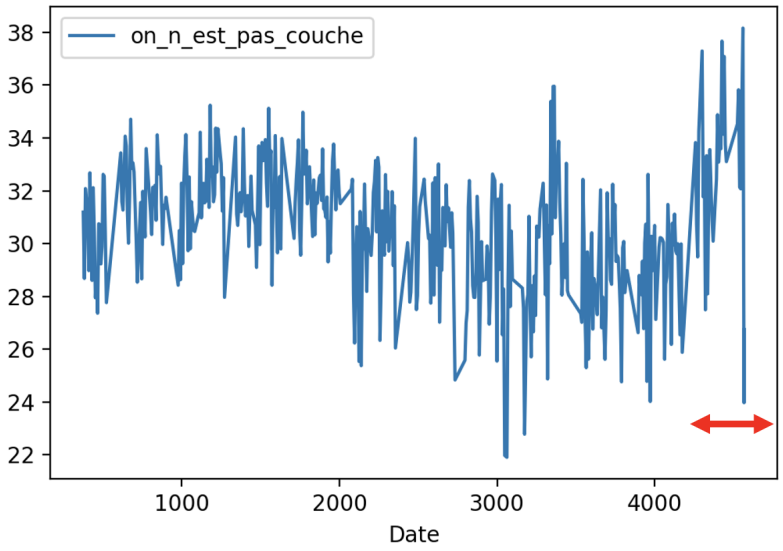

Рис. 3.2: Зависимость средней длины фразы для передачи On n'est pas couché от времени(кол-во дней с 01.01.2008)

On n'est pas couché. В 2019 году еженедельное ток-шоу Лорана Рукье On n'est pas couché, транслируемое на France 2, отказывается от модели, которая принесла ему известность. До этого, каждый сезон наблюдалось противостояние двух, часто спорных, обозревателей, способных спровоцировать жаркие дебаты. Известны, в частности, споры между Наташей Полони и Кароном Аймери за период 2012-2014гг [2]. В начале 2019 года каждую неделю в этой передаче видим двух новых обозревателей[9], для которых характерно более спокойное ведение диалога.

Данная политика телепередачи оказывает значительное влияние на продолжительность сегментов, которые, как можно заметить, значительно дольше в этом сезоне, отмеченном красной стрелкой на графике. Также в сезоне 2020 года выпуски On n'est pas couché реже появляются в виде выдержек в социальных сетях, и, как следствие, аудитория

канала уменьшилась. Следует отметить, что это ток-шоу было заменено новым, представленным Лораном Рукье. Здесь можно отметить аргумент в пользу гипотезы, упомянутой во введении: дебаты, выраженные короткими эмоциональными обменами, склонны привлекать аудиторию и могут быть удачной стратегией для каналов.

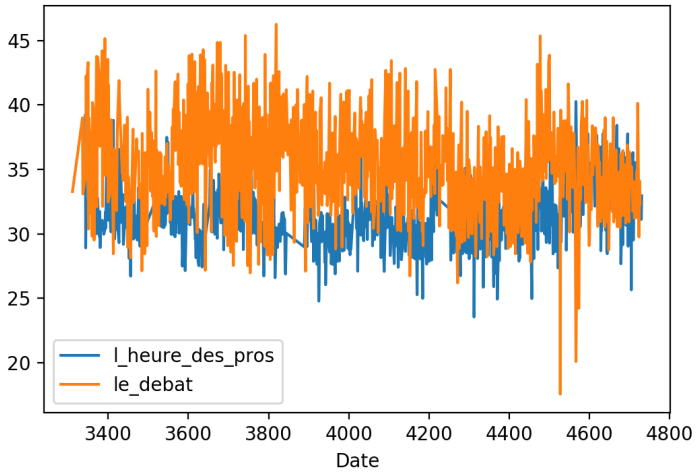

Рис. 3.3: Зависимость средней длины фразы для передач L'heure des pros и le débat от времени(кол-во дней с 01.01.2008)

LCI vs. CNEWS. Программы le débat, организованная Арлетт Чабот(Arlette Chabot), и L'heure des pros, проводимая Паскалем Про(Pascal Praud), были запущены в 2016 году. Заметно, что среднее количество слов по выпуску значительно ниже для трансляции CNEWS (31.43) чем у LCI (35.29). Гипотеза о том, что дебаты являются более оживленными в программе Паскаля Про подтвердилась и здесь. В настоящее время нередко можно увидеть в социальных сетях выдержки из жарких дебатов между Паскалем Про, обозревателями и гостями, и эти выдержки характеризуются короткими фразами [5]. В передаче LCI разговоры спокойнее. Графически эта разница отчётливо видна.

CNEWS : Неоднородная стратегия. После рассмотрения предыдущих наблюдений возникает вопрос: можно ли обнаружить паттерны ведения шоу Паскаля Про на других программах канала? Сравнение программ L'heure des pros и Face и l'info показывает более высокое среднее значение длины фразы для последнего (35.58 слова на сегмент), того же порядка длины, что наблюдалось в Le Débat. Наиболее вероятным предположением, объясняющим это различие, является формат двух программ. Полемические выдержки из этих передач отличаются: если для шоу Паскаля Про характерны многочисленные смены спикеров во время передачи, то в передаче Кристин Келли больше заметны моменты, в которые Эрик Земмур имеет возможность произносить провокационную речь. У него

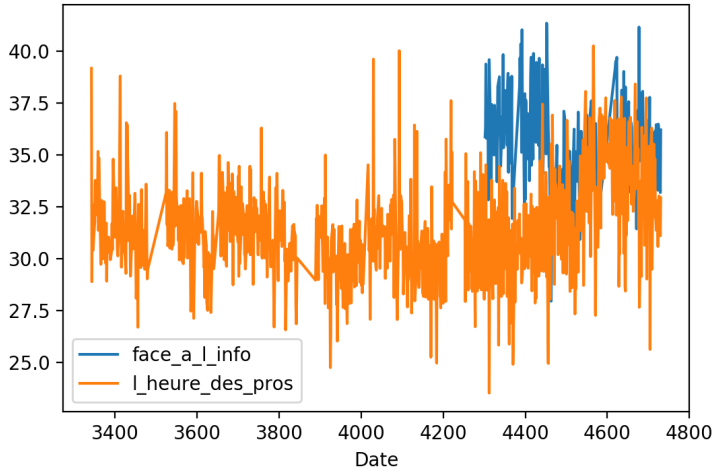

Рис. 3.4: Сравнение передач L'heure des pros (CNEWS) и Face l'info (CNEWS) по продолжительности речи от времени(кол-во дней с 01.01.2008)

есть время, чтобы раскрыть свои теории, иногда, без какого-либо перебивания со стороны. Можно заключить, что индикатор, который мы использовали здесь, заслуживает того, чтобы считаться удачным.

Рассуждения о длине фраз показали много интересных моментов из телепередач и позволили обнаружить значительные различия между выпусками. Также иногда такой анализ позволяет отметить четкую эволюцию политики телеканала с течением времени. Длина фразы является наблюдаемым, не требующим предварительных операций. В то же время, этот показатель не позволяет провести более глубокий анализ дебатов, который мы стремимся провести.

3.2 Анализ тональности

Оценка тональности дебатов – это присвоение каждому сегменту коэффициента, который характеризует его эмоциональную окраску. Данный подход является основой NLP-метода, называемого сентимент анализ[11][12].

В нашей работе этот коэффициент может принимать все значения между -1 (очень негативное восприятие) и 1 (очень позитивное восприятие).

Метод

В данной работе используется метод TextBlob, так как он подходит для анализа текстов на французском языке. На нескольких примерах мы наблюдаем следующие коэф-

фициенты, рассчитанные при помощи данной библиотеки:

Сегмент	**Тональность**
cet homme est mauvais je le hais il doit souffrir et mourir	-0.71
le ciel est gris	-0.18
le ciel est bleu	0.04
j'ai hâte de partir en vacances le paysage est beau et le soleil est bon	0.75

Таблица 3.2: Пример подсчёта тональностей для нескольких предложений.

Библиотека Textblob предоставляет различные способы анализировать тональность высказываний, при использовании соответствующего параметра analyzer. Этот параметр отвечает за то, на вход какой модели будет подаваться текст. Соответственно, необходимо выбирать analyzer в зависимости от задачи. Для текстов на французском языке доступен PatternAnalyzer. Эта модель основана на заранее написанных правилах. Кроме недостатка, заключающегося в больших трудозатратах на создание данной модели, можно отметить высокую зависимость от качества перевода. Например, для английской версии существуют примеры очевидно неправильной работы данного анализатора, несмотря на то, что английская версия является одной из самых проработанных да и сам язык считается не самым сложным языком для NLP, в отличие от русского или французского.

Также в библиотеке Textblob доступен анализатор для оценки тональности фразы, основанный на наивном байесовском классификаторе NaiveBayesAnalyzer.

Модуль тональности

Первоначально мы стремимся наблюдать хронологическую эволюцию тональности для каждого выпуска. Поэтому мы агрегируем все сегменты в одном файле, а затем переходим к расчету тональности для каждого из них. Для шоу Face à l'info график тональностей по всем сегментам не очень читаем, так как почти 100'000 сегментов должны быть отображены, и вследствие большого количества сегментов экстремальные значения регулярно принимаются. Необходимо найти способ анализировать эти данные. Чтобы сгладить кривую, агрегируем сегменты по пакетам и извлекаем среднюю тональность для каждого пакета.

Для других выпусков получены следующие графики:

Из построенных графиков можно сделать вывод, что вариации минимальны (порядка нескольких сотых) и значения сосредоточены вокруг одного среднего. Этим резуль-

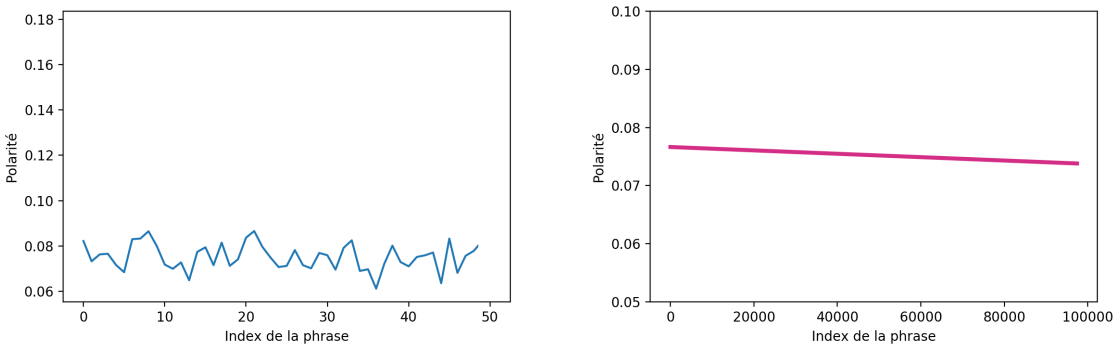

Рис. 3.5: Хронологическая тональность сегментов *Face à l'info* после агрегации по пакетам и регрессионная линия по всем сегментам.

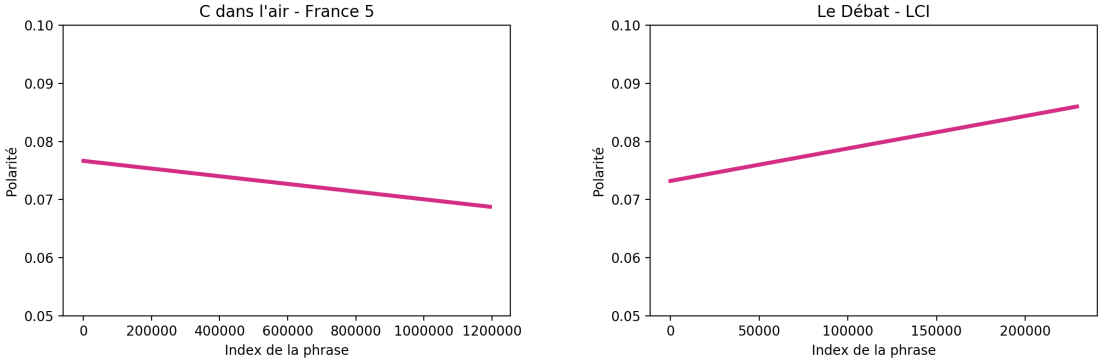

Рис. 3.6: Линейная регрессия тональностей сегментов для выпусков *C dans l'air* et *Le Débat*.

татам можно дать следующие объяснения:

- подавляющее большинство сегментов имеют нулевую поляризацию, и они ведут к среднему, близкому к 0
- позитивные и негативные сегменты нейтрализуют друг друга

Несмотря на некоторое сходство тенденций в графиках средней длины фраз и средних тональностей для передачи Face à l'info(уменьшение средней величины во второй половине графика), текущие наблюдения не позволяют нам делать адекватные выводы. Нам необходимо усовершенствовать этот индикатор, фильтруя значения, сконцентрированные вокруг среднего, и разделяя позитивные и негативные тональности сегментов. Кроме того нужно помнить, что позитивная тональность может быть переоценена, потому что среднее, вероятно, подтягивается рекламой. Мы должны предположить, что передачи содержат рекламные части сопоставимой продолжительности, чтобы рассматривать его влияние как одинаковое на всех каналах. Это предположение не является действительным для госу-

дарственных телеканалов, на которых реклама ограничена.

Фильтрация шумов

Далее необходимо провести улучшения нашего индикатора, чтобы получить больше информации об определенных характеристиках телепередач. Кажется целесообразным анализировать позитивные и негативные сегменты отдельно, как и предлагалось ранее. Мы сохраняем сегменты, абсолютное значение тональности которых больше, чем 0.2. Таким образом, идея состоит в том, чтобы сосредоточить внимание на весьма поляризованном фрагменте полилогов. Затем, мы отслеживаем эволюцию значений тональностей по модулю и частоте. Временная эволюция, как можно заметить, очень мала, даже при разде-

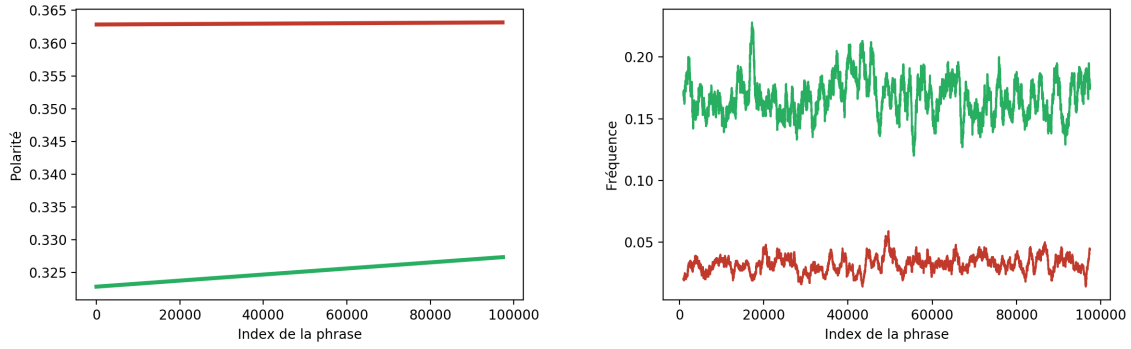

Рис. 3.7: Слева, график линейной регрессии (абсолютное значение), полученные для *Face à l'info*, позитивные тональности в зеленом цвете, негативные – в красном. Справа скользящее среднее значение частоты появления сильно позитивной/негативной тональности по тысяче сегментов.

лении позитивных и негативных тональностей. Чтобы получить более сильные вариации, необходимо ограничить фильтр более экстремальными значениями (например, абсолютная тональность больше, чем 0.5), но мы тогда работаем над слишком малым количеством сегментов. Тем не менее, сделаны выводы, которые позволяют нам понять предыдущие результаты:

- в абсолютном значении негативные предложения кажутся более выраженными
- по частоте позитивные предложения более многочисленны

Сочетание этих двух факторов во многом объясняет наблюдаемую выше нейтрализацию тональностей при анализе эмоциональной составляющей полилогов. Аналогичные результаты можно наблюдать и по другим предоставленным передачам.

тональность для каждого спикера Еще одной информацией, которая может быть использована для изучения тональности выпусков, является идентификация спикеров. Транскрипции INA включают информацию об участвующих сторонах полилога, каждой из которых был присвоен код для его идентификации. Далее рассмотрим тональности на основе этих данных, чтобы увидеть, оказывают ли определённые спикеры влияние на общую тональность программы.

Рассмотрим тональности для каждого спикера в хронологическом порядке для программы *L'Heure des pros*.

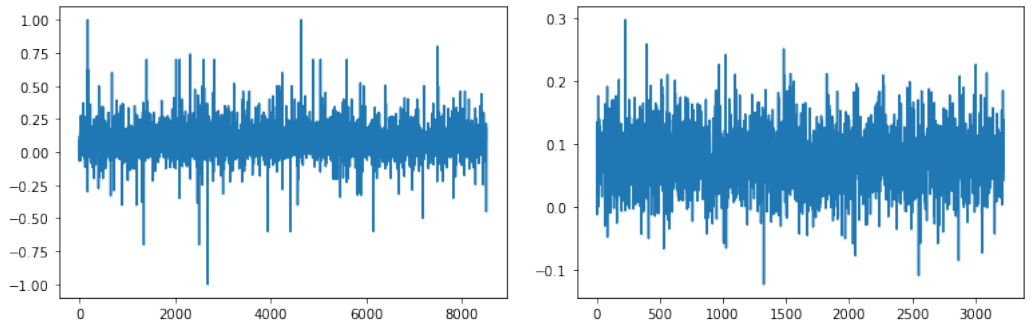

Рис. 3.8: Тональность для каждого спикера передачи *L'Heure des pros* без фильтров (слева) и с фильтрацией с минимальным количеством речей равным 10 (справа).

На нефильтрованном графике видно, что многие участники имеют большие по модулю тональности. Большое количество ораторов приводит нас к выводу, что некоторые из гостей записывались несколько раз под разными индексами в транскрипции ИНА. Поэтому был сделан выбор сохранять только тех, участников, которые говорили более десяти раз за передачу. Результат является менее значительным, и нельзя сделать вывод, что тональность программы обусловлена поведением определенного участника. Действительно, участники со значительным количеством реплик не позволяют сделать выпусках, несмотря на то, что у метода достаточно понятная основа. Можно сделать вывод, что сентимент анализ недостаточно хорошо улавливает настроения на несколько зашумленных данных.

Главные проблемы метода

Сентиментальный анализ является теоретически мощным инструментом, который предлагалось использовать в качестве основы для нашего индикатора. Однако присвоение тональности сегментам не позволило выделить необходимые характеристики для каждой программы. Все они, следуют одному же паттерну, и необходимая информация не бы-

ла замечена либо исчезла в потоке нейтральных сегментов. Фильтрации не позволили избавиться от недостатков без утери информации, и объединение тональностей в соответствии с участвующими сторонами не справляется с проблемами, налагаемыми качеством и форматом транскрипций. Таким образом, этот индикатор интересен с точки зрения полученных выводов и имеет ограниченное использование для конкретных случаев.

Глава 4

Заключение

Исследование основывалось на изучении метаданных и использовании и совершенствовании методов, основанных на сентимент анализе. Другой частью поставленной задачи являлось сравнение данных выпусков нескольких форматов полилогов(ток-шоу с различными экспертами, интервью) с целью тестирования индикаторов длины фраз и их тональности для самых подходящих для этого выпусков. И на основе этих тестов сделаны выводы, совпадающие с мнением экспертов в области анализа теледебатов. Таким образом в качестве источников данных были использованы транскрипции телепередач, предоставленные Национальным институтом(INA). Проведён анализ метаданных, в ходе которого были выявлены изменения в политике ведения телепередач. Также был выработан индикатор, основанный на анализе тональности. Были продемонстрированы основные недостатки метода, например, высокая чувствительность к чистоте транскрипций.

Выражается благодарность Этьену Олиону за содействие в выполнении работы и помощь в получении данных. Также выражается благодарность INA за предоставление данных в удобном формате в сжатые сроки.

Список литературы

[1] *Chaînes d'info : l'extrême droite en croisière*. Acrimed. URL: https://www.acrimed.org/Chaines-d-info-l-extreme-droite-en-croisiere?var_mode=calcul.

[2] *Clash Caron-Polony : pourquoi le divorce était inéluctable*. Le Figaro. URL: https://www.lefigaro.fr/vox/medias/2014/03/10/31008-20140310ARTFIG00096-clash-aymeric-caron-natacha-polony-pourquoi-le-divorce-etait-ineluctable.php.

[3] *Des questions jamais entendues. Crise et renouvellements du journalisme politique à la télévision*. Erik Neveu, Politix. URL: https://www.persee.fr/doc/polix_0295-2319_1997_num_10_37_1648.

[4] *Discrimination of Blacks by Police*. URL: https://www.pnas.org/content/114/25/6521.full.

[5] *L'affrontement entre Pascal Praud et Claire Nouvian en trois actes*. L'Express. URL: https://www.lexpress.fr/actualite/medias/la-polemique-entre-pascal-praud-et-claire-nouvian-en-trois-actes_2077157.html.

[6] *L'ANALYSE DU CONTENU DES DÉBATS POLITIQUES TÉLÉVISÉS*. Gilles Gauthier. URL: http://documents.irevues.inist.fr/bitstream/handle/2042/15229/?sequence=1.

[7] *Les bienséances de l'échange politique. Naissance d'une tribune politique télévisuelle*. Éric Darras, Politix. URL: https://www.persee.fr/doc/polix_0295-2319_1997_num_10_37_1647.

[8] *Loi du 5 mars 2009 relative à la communication audiovisuelle et au nouveau service public de la télévision*. URL: https://www.vie-publique.fr/loi/20544-television-service-public-publicite-chaines-publiques-audiovisuel.

[9] *On n'est pas couché : les nouveaux chroniqueurs de l'émission dévoilés*. L'Express. URL: https://www.lexpress.fr/culture/tele/on-n-est-pas-couche-les-nouveaux-chroniqueurs-de-l-emission-devoiles_2095936.html.

[10] *Учебник по Python Pandas*. CoderLessons. URL: https://coderlessons.com/tutorials/python-technologies/vyuchit-python-panda/uchebnik-po-python-pandas.

[11] *Sentiment Analysis: A Definitive Guide*. MonkeyLearn. URL: https://monkeylearn.com/sentiment-analysis/.

[12] *Understanding Political Twitter*. Towards Data Science. URL: https://towardsdatascience.com/understanding-political-twitter-ce3476a38377.

[13] *User Guide*. pandas. URL: https://pandas.pydata.org/docs/user_guide/index.html.