СОДЕРЖАНИЕ

Введение 3
Глава 1. Аналитический обзор 4
1.1. Понятие тонального анализа 4
1.2. Исследование существующих методов анализа тональности текстовых сообщений 5
1.3. Исследование Tomita-парсера 7
1.4. Исследование тональных словарей 9
1.5. Выбор способа реализации 11
Глава 2. Разработка прототипа 12
2.1. Краткое описание алгоритма 12
2.2. Разработка правил для Tomita-парсера 12
2.3. Присвоение тональности атрибутам 17
2.4. Разработка графического интерфейса 18
2.5. Тестирование разработанного прототипа 19
Заключение 26
Список использованных источников 27
Приложение 1. Программный код скриптов на C# 28

ВВЕДЕНИЕ

На данный момент информация хранится преимущественно в электронном виде. Обработка текстовых сообщений представляет собой большой интерес в различных сферах. Например, отзывы потребителей товаров или услуг в Интернете позволяют производителям улучшить качество предоставляемой продукции, выявить потребности потребителей, а потребителям сделать выбор в пользу того или иного товара или услуги. Анализ комментариев в социальных помогает отслеживать и удалять комментарии, содержащие в себе оскорбления, призывы к противоправным действиям. Анализ текстовых сообщений используется во многих сферах науки и бизнеса. Текстовые сообщения представляют собой текст на естественном языке, так людям привычно воспроизводить и воспринимать информацию, но для обрабатывать большие объемы данных, написанных таким способом, затруднительно.

Для извлечения структурированных данных из текстовых сообщений, написанных на естественном языке, существует большое количество инструментов (в особенности, для английского языка). Для русского языка существует инструмент от компании Яндекс - "Томита-парсер". Для обработки полученной структурированной информации используются методы компьютерной лингвистики. Одной из задач компьютерной лингвистики является анализ тональности текста. Тональность текста - это эмоциональное отношение автора текста к какому-то объекту. Тональность часто определяется как "позитивная"или "негативная также есть другие оценки. Методами определения тональности могут быть основаны на использовании тональных словарей или построены на машинном обучении.

Цель исследования — разработка прототипа с применением Tomita-парсера для анализа атрибутивной тональности текстовых сообщений. В соответствии с целью исследования, определяются следующие задачи работы:

A. Исследовать возможности Tomita-парсера для извлечения тональной информации из текстовых сообщений.
B. Разработать правила для Томита-парсера для извлечения "эмоционально–окрашенных"слов из текстовых сообщений.
C. Разработать модуль на C#, извлекающий тональные слова из текстовых сообщений.
D. Тестирование разработанного модуля.

ГЛАВА 1. АНАЛИТИЧЕСКИЙ ОБЗОР

В данной главе будет проведено исследование понятия «тональный анализ»1.1 и его методов 1.2, проведено исследование основных возможностей Tomaita-парсера 2.2, проведено исследование доступных тональных словари и сделан выбор в пользу одного из них 2.2, а также выбран способ создания прототипа для анализа текстовых сообщений 1.5.

1.1. Понятие тонального анализа

Тональный анализ (англ. Sentiment analysis) — это область компьютерной лингвистики, которая занимается изучением мнений и эмоций в текстовых документах [7]. Тональность текста в целом определяется лексической тональностью составляющих его единиц и правилами их сочетания. Присвоение тональности может происходить по разным правилам:

- классификация по бинарной шкале (оценка может быть либо позитивной, либо негативной)
- классификация по многополосной шкале (дискретные значения из ограниченного промежутка)
- системы шкалирования (непрерывная шкала оценок, отрезкам на которой соответствуют словесные описания тональности)

В данной работе будет использоваться система шкалирования, так как она позволяет оценить тональность не округляя значение, полученное с помощью вычислений. Также в работе будут использоваться 4 варианта словесной оценки тональности:

- позитивная
- негативная
- нейтральная
- противоречивая

1.2. Исследование существующих методов анализа тональности текстовых сообщений

Так как запрос на тональный анализ текста актуален на данный момент [8], то существуют различные методы для его проведения. Каждый из способов имеет свои преимущества и недостатки, которые будут подробно описаны далее.

A. *Методы, основанные на машинном обучении*

Данные методы используют с ранее размеченные эталонные блоки и относят их к негативу или позитиву на основании полученного результата сравнения [3]. При таком подходе требуется большой объем вводных данных. Преимуществами данных методов в гибкости, так как система без труда переобучается, и в высоких показателях полноты извлечения информации [5]. Недостатками же являются [4]:

- трудоемкость подготовки и обучения (требуется большой объем текстов)
- сложность выявления и исправления ошибки

B. *Методы, основанные на правилах и словарях*

Данные методы основываются на заранее составленных тональных словарях и правилах с использованием лингвистического анализа [2].Тональные словари должны содержать в себе достаточное количество слова для использования в исследуемой предметной области. Правила могут представлять из себя регулярные выражения или основываться на связи слов в тексте. Чтобы проанализировать текстовое сообщение методом, основанном на правилах и словаре, необходимо сначала составить тональный словарь, затем присвоить всем словам тональность и после сделать вывод об общей тональности текста. Общая тональность может определяться как среднее арифметическое всех тональных значений. Преимуществом данных методов является точность. Среди недостатоков можно отметить следующие:

- трудоемкость составления словаря, достаточно полным для исследования нужной предметной области
- неоднозначность присвоения тональности (слово «громкий» может иметь положительную окраску, когда речь идет о громком голосе певицы, но негативную, когда речь идет о шуме)

- сложности определения тональности при наличии орфографических ошибок в тексте

Для работы был выбран метод, основанный на тональном словаре, но алгоритм присвоения тональности при этом был изменен. Данному методу было отдано предпочтение по следующим причинам:

- результат не зависит от используемых текстов
- высокая точность
- не требуется большой объем данных для получения удовлетворительного результата

Как было упомянуто ранее, алгоритм присвоения тональности подвергся модификации. Стандартно словам в текстовом сообщении присваивается определенная тональность, полученная при использовании тонального словаря, а затем путем арифметических операций получается некоторое среднее значение,по которому будет сделан вывод о тональности всего текста. При таком алгоритме высока вероятность в противоречивом тексте получить вывод, свидетельствующий об отсутвии тональности («нейтральная» тональность). Например, «Сегодня прекрасная погода, но ужасное настроение». Согласно первоначальному алгоритму слову «прекрасная» будет присвоена позитивная тональность, а слову «ужасное» негативная. Будет ли значение тональности этих в цифровом эквиваленте ровно противоположно сказать трудно, но скорее всего их разность все же будет близка к нулю. Таким образом, велик риск, что данному предложению будет присвоена нейтральная тональность, хотя автор дал яркую эмоционально оценку словам «погода» и «настроение». Чтобы избежать подобного результата можно анализировать текст, присваивая общую тональность атрибутам. Тогда в результате будет получено следующее: «настроение» - негативная тональность, «погода» - позитивная тональность. А общая тональность текста будет скорее противоречивой. Данный подход позволяет более точно оценить тональность текста и не проигнорировать эмоциональное отношение автора к объектам в этом тексте.

1.3. Исследование Tomita-парсера

Tomita-парсер - это инструмент, разработанный компанией Яндекс, позволяющий извлекать структурированные данные из текста на естественном языке. Вычленение фактов происходит при помощи контекстно-свободных грамматик и словарей ключевых слов. Парсер позволяет писать свои грамматики и добавлять словари для нужного языка [10]. Парсер был назван в честь японского ученого Масару Томита, который разработал алгоритм для анализа текста на естественном языке. Именно этот алгоритм лежит в основе работы Томита-парсера. С помощью парсера можно выделять цепочки слов или факты по написанным пользователем шаблонам [11]. Парсер включает в себя три стандартных лингвистических процессора: токенизатор (разбиение на слова), сегментатор (разбиение на предложения) и морфологический анализатор (mystem). Файлы, используемые для работы с парсером приведены на рис.1.1.

Содержание	Формат	Примечания
config.proto — конфигурационный файл парсера. Сообщает парсеру, где искать все остальные файлы, как их интерпретировать и что делать.	Protobuf	Нужен всегда.
dic.gzt — корневой словарь. Содержит перечень всех используемых в проекте словарей и грамматик.	Protobuf / Gazetteer	Нужен всегда.
mygram.cxx — грамматика	Язык описания грамматик	Нужен, если в проекте используются грамматики. Таких файлов может быть несколько.
facttypes.proto — описание типов фактов	Protobuf	Нужен, если в проекте порождаются факты. Парсер запустится без него, но фактов не будет.
kwtypes.proto — описания типов ключевых слов	Protobuf	Нужен, если в проекте создаются новые типы ключевых слов.

Рис.1.1. Файлы для работы с Томита-парсером

Согласно информации на рис.1.1 минимальный набор файлов для работы с парсером - файл конфигурации и корневой словарь. После запуска парсера формируется как минимум один файл PrettyOutput.html, который в табличном виде содержит в себе цепочки, которые выделила грамматика. На рис.1.2 пример вывода полученных парсером цепочек. Грамматики парсера работают следующим образом: предложения представляют собой цепочки, из них выделяются подцепочки, а подцепочки интерпретируются в разбитые по полям факты.

Чтобы извлечь из текста подцепочки, представляющие собой существительные, нужно записать следующее правило грамматики:

$$S- > Noun; \tag{1.1}$$

С помощью этого правила из уже знакомого предложения «Сегодня прекрасная погода, но ужасное настроение» будут извлечены подцепочки в нормализованном виде: «сегодня», «погода» и «настроение».

Сегодня прекрасная погода , но ужасное настроение **EOS**

Text	Type
сегодня	TAuxDicArticle [грамматика]
погода	TAuxDicArticle [грамматика]
настроение	TAuxDicArticle [грамматика]

Рис.1.2. Цепочки, полученные в результате работы парсера

С помощью Томиты-парсера становится возможным извлечение из текстовых сообщений атрибутов и тональных слов, к ним относящихся.

1.4. Исследование тональных словарей

Так как для исследования был выбран метод, основанный на работе с тональными словарями, то необходимо было выбрать или составить самостоятельно словарь для дальнейшего использования. Необходимо определить требования к словарю, чтобы среди обилия вариантов выбрать наиболее подходящий вариант. Для определения тональности атрибутов было решено использовать прилагательные, относящиеся к этим атрибутам. Прилагательное - это часть речи, обозначающая качество, свойство или принадлежность предмета и изменяющаяся по падежам, числам и родам. Именно с помощью прилагательных в большинстве случаев дается эмоциональная оценка объектам.

По этой причине в тональном словаре должен быть широкий набор общеупотребительных прилагательных и существительных. Для исследования не была выбрана узкая область, предполагается, что работа будет происходить с текстами, содержащими в себе бытовую лексику. Для интерпретации оценки тональности ранее была выбрана методика, основанная на системе шкалирования, поэтому необходимо, чтобы словарь содержал в себе числовую оценку тональности из некоторого непрерывного диапазона.

После продолжительного поиска было найдено 3 возможных варианта тональных словарей для дальнейшего использования:

- тональный словарь PolSentiLex. Достоинства: имеет большую базу слов, проект реализован специалистами Лаборатории интернет-исследований НИУ ВШЭ – СПб. Недостатки: имеет дискретную систему оценки тональности (-2, -1, 0, 1, 2).
- словарь оценочных слов и выражений русского языка РуСентиЛекс. Достоинства:содержит в себе фразы и выражения, а также сленг, всего в базе более 12 тысяч слов и выражений. Недостатки: тональность определяется словесно (позитивная, негативная, нейтральная или неопределенная оценка)
- Тональный словарь русского языка КартаСловСент. Достоинства: включает в себя слова и выражения русского языка, снабжённые тональной меткой («положительное», «отрицательное», «нейтральное») и скалярным значением силы эмоционально-оценочного заряда из непрерывного диапазона [-1, 1], содержит в себе более 46 тысяч записей. Недостатки: для данной задачи не обнаружено.

Наиболее подходящим для данных задач словарем оказался словарь КартаСловСент. Фрагмент файла словаря можно увидеть на рис.1.3.

```
term;tag;value;pstv;ngtv;neut;dunno;pstvNgtvDisagreementRatio
абажур;NEUT;0.08;0.185;0.037;0.58;0.198;0.0
аббатство;NEUT;0.1;0.192;0.038;0.578;0.192;0.0
аббревиатура;NEUT;0.08;0.196;0.0;0.63;0.174;0.0
абзац;NEUT;0.0;0.137;0.0;0.706;0.157;0.0
абиссинец;NEUT;0.28;0.151;0.113;0.245;0.491;0.19
абитуриент;NEUT;0.23;0.235;0.049;0.5;0.216;0.0
абитуриентка;NEUT;0.34;0.294;0.029;0.461;0.216;0.0
абонемент;NEUT;0.18;0.232;0.056;0.56;0.152;0.0
абонементный;NEUT;0.19;0.21;0.07;0.5;0.22;0.0
абонент;NEUT;0.0;0.075;0.075;0.675;0.175;0.0
абонентный;NEUT;0.0;0.073;0.0;0.61;0.317;0.0
абонентский;NEUT;0.0;0.071;0.0;0.822;0.107;0.0
абордаж;NGTV;-0.45;0.07;0.33;0.38;0.22;0.0
абордажный;NEUT;-0.19;0.137;0.171;0.384;0.308;0.11
абориген;NEUT;-0.02;0.063;0.167;0.624;0.146;0.0
```

Рис.1.3. Тональный словарь русского языка КартаСловСент

В первой строчке рис.1.3 содержатся названия полей:

- term - слово или словосочетание
- tag - метка тональности: PSTV («положительное»), NGTV («отрицательное»), NEUT («нейтральное»)
- value - скалярное значение эмоционально-оценочного заряда из непрерывного диапазона [-1, 1], где +1 соответствует входам с максимально положительной окраской, -1 — входам с максимально отрицательной окраской, 0 — входам с нейтральной окраской (то же, что отсутствие окраски)
- а также другие поля, характеризующие доли голосов за определенную тональности

В данном словаре уже используется система для перевода числового значения тональности в словесный: $[-1, -0.3]$ - негативная тональность, $[0.5, 1]$ - позитивная тональность, остальное - нейтральная. Эта система будет использоваться в дальнейшем для всего приложения.

1.5. Выбор способа реализации

Для запуска Томиты-парсера используется командная строка, а для просмотра полученных результатов html файл, хоть в командной строке после завершения работы и появляется строка xml формата с этими же результатами. Это сделано потому, что воспринимать информацию человеку удобнее, когда она визуализирована, например, в виде таблицы. По этой причине разрабатываемый прототип должен иметь графический интерфейс и простой дизайн. Для создания приложений с графическим интерфейсом отлично подходит технология Windows Forms. Разработка приложения будет производиться с использованием языка C#.

ГЛАВА 2. РАЗРАБОТКА ПРОТОТИПА

В данной главе последовательно описывается разработка прототипа для тонального анализа текста. В первом параграфе главы 2.1 описан краткий алгоритм работы прототипа. В параграфе 2.2 подробно описан процесс разработки правил грамматики для Томита-парсера, а также содержание других файлов для работы с парсером. В параграфе 2.3 описан процесс присвоения тональности извлеченным парсером словам и формирование окончательного результата по анализу атрибутивной тональности текста. Параграф 2.4 содержит в себе описание внешнего вида приложения, а точнее - описание графического интерфейса. Тестирование разработанного приложения производится в параграфе 2.5 на основе сравнения ожидаемых результатов и полученных.

2.1. Краткое описание алгоритма

После запуска приложения пользователь вводит текст. Далее этот текст обрабатывается приложением в несколько этапов, которые представлены в коде приложения методами:

- метод GetText() удаляет старый файл с текстом (если есть), записывает данные введенные от пользователя в новый файл с названием test.txt
- метод makeCodeTomita(int mode) в зависимости от аргумента формирует first.cxx — файл грамматики двумя разными способами
- метод StartTomita() осуществляет запуск парсера
- метод ReadFacts() отвечает за получение фактов из парсера и их обработку
- метод UpdateGrid() формирует вывод результатов обработки фактов пользователю

2.2. Разработка правил для Tomita-парсера

В первой главе были изучены основные принципы работы с Томита-парсером. Как уже упоминалось в параграфе ,для определения тональности необходимо получить атрибуты и прилагательные к ним относящиеся. Атрибутами будут выступать существительные. Таким образом, правила для парсера должны извлекать цепочки, состоящие из существительного и прилагательного (прилагательных).

Файл грамматики был создан с именем first.cxx. Для того, чтобы парсер знал об этом файле, в файле словаря необходимо написать следующую команду (см. рис.2.1).

```
TAuxDicArticle "грамматика"
{
    key = { "tomita:first.cxx" type=CUSTOM }
}
```

Рис.2.1. mydic.gzt

В конфигурационном файле config.proto нужно добавить грамматику (см. рис.2.2).

```
Articles = [
{ Name = "грамматика" } ]
```

Рис.2.2. config.proto

Далее необходимо написать правила для Томиты-парсера. У атрибута может быть одно существительное, а может быть несколько разделенных запятыми и/или союзами. Извлечение фактов будет проходить в два этапа: сначала нужно извлекать цепочку, содержащую в себе атрибут и прилагательные, а затем из группы прилагательных выделять одиночные слова.

Правила можно прописать в одном файле, но тогда есть риск, что некоторые группы прилагательных будут разбиваться неверно. Например, если на вход подается текст «мой рабочий костюм» , то сначала выделится атрибут «костюм» и группа слова «мой рабочий» . Затем, парсер будет запущен снова и несмотря на наличие правил для извлечения только прилагательных, парсер извлечет атрибут «рабочий» , так как в данном словосочетании, действительно, складывается впечатление, что «рабочий» - это существительное. Для того, чтобы этого избежать, было принято решение формировать файл грамматики first.cxx во время работы программы. В таком случае сначала будут извлечены все цепочки содержащие атрибуты, затем правила в грамматике поменяются и на вход парсеру будут подаваться группы прилагательных, которые будут разбиваться по одному по новым правилам. Для извлечения цепочек существительное + прилагательные были написаны следующие команды (см. рис.2.3)

В строке 1 (рис.2.3) формируется запись, представляющая собой одно прилагательное (Adj - прилагательное, причастие, порядковое числительное или

```
1 AdjCoord->Adj;
2 AdjCoord->AdjCoord <gnc-agr[1], rt> ',' AdjCoord <gnc-agr[1]>;
3 AdjCoord->AdjCoord <gnc-agr[1], rt> 'и' AdjCoord <gnc-agr[1]>;
4 S->Noun<gnc-agr[1]> interp(Fact.Noun) AdjCoord<gnc-agr[1]>+
    interp(Fact.Adj::norm = "m,sg");
5 S->AdjCoord <gnc-agr[1]> +interp(Fact.Adj::norm = "m,sg") Noun
     <gnc-agr[1]> interp(Fact.Noun);
```

Рис.2.3. first.cxx

```
1 S->Adj interp(FactAdj.A::norm = "m,sg");
```

Рис.2.4. first.cxx

местоименное прилагательное). В строке 2 (рис.2.3) обрабатывается случай, когда прилагательные разделены запятыми. <gnc-agr[1]> указывает на согласование по роду, числу и падежу. В строке 3 (рис.2.3) аналогичным образом обрабатывается случай, когда прилагательные разделены союзом «и». В четвертой строке S является вершиной формируемой цепочки. Цепочка будет формироваться так: существительное Noun, согласованное по роду, числу и падежу с одной или более AdjCoord записью. На количество «один или более» указывает «+» после AdjCoord. Фрагмент «interp(Fact.Adj::norm = «m,sg»)» указывает на извлечение слова, которое будет нормализовано: мужской род («m»), единственное число («sg»). Нормализация нужна для поиска по словарю, в котором слова представлены в единственном числе, в мужском роде. Команда «interp» позволяет интерпретировать факты, то есть распределить подцепочки из выделенной грамматикой цепочки по полям факта «Adj». В строке 5 (рис.2.3) обрабатывается случай, если прилагательные стоят перед существительным.

Для извлечения из текста только прилагательных в файле first.cxx нужно сформировать следующую команду:

Команда (рис.2.4) извлекает прилагательные и нормализует их по уже описанным правилам.

Также необходимо создать файл facttypes.proto и записать туда информацию об извлекаемых фактах:

В строках 1 и 6 (рис.2.5) задаются имена фактам. В строках 3, 4 и 8 указывается, что поля фактов это строки с именами Noun, Adj, A (эти имена указываются в интерпретации в файле грамматики first.cxx (рис.2.3 , рис.2.4) {Имя факта}.{Имя поля}).

```
message Fact: NFactType.TFact
{
    required string Noun = 1;
    required string Adj = 2;
}
message FactAdj: NFactType.TFact
{
    required string A = 1;
}
```

Рис.2.5. facttypes.proto

```
Facts = [
    { Name = "Fact" },
    { Name = "FactAdj" }
]
```

Рис.2.6. config.proto

Для того, чтобы было удобнее программно получать информацию от парсера, можно формировать файл output.txt. Чтобы сформировать данный файл, необходимо в config.proto добавить следующий фрагмент кода (рис.2.6):

Данный код (рис.2.6) указывает на сформированные ранее правила интерпретации цепочек.

Текст, получаемый парсером на обработку, хранится в файле test.txt и представляет собой текстовое сообщение, записанное на естественном языке. Пример извлечения фактов парсером из текста «Сегодня прекрасная погода, но ужасное и мерзкое настроение» см. на рис.2.7 с применением грамматики (рис.2.3).

```
Сегодня прекрасная погода , но ужасное и мерзкое настроение
        Fact
        {
                Noun = погода
                Adj = прекрасный
        }
        Fact
        {
                Noun = настроение
                Adj = ужасный и мерзкий
        }
```

Рис.2.7. Файл output.txt

Были извлечены атрибуты «погода» и «настроение», так как они имели согласованные с ними прилагательные. Теперь согласно алгоритму на вход парсеру

пойдут поочередно строки «прекрасный» и «ужасный и мерзкий». Первую строку проверять нет необходимости (понятно, что это одно слово и оно уже может использоваться для поиска в тональном словаре), чтобы сообщить это программе будет введена проверка на наличие пробела в строке (как минимум один пробел между прилагательными будет, если прилагательных 2 или более). Результат анализа парсером строки «ужасный и мерзкий» см. на рис.2.8.

```
ужасный и мерзкий
        FactAdj
        {
                A = ужасный
        }
        FactAdj
        {
                A = мерзкий
        }
```

Рис.2.8. Файл output.txt

Прилагательные были получены верно в нормализованном виде, значит теперь можно переходить к этапу назначения тональности атрибутам. Для того, чтобы запускать парсер из приложения был разработан метод StartTomita(), который представлен на листинге (рис.2.9). Данный метод позволяет открыть командную строку без создания окна, перейти в нужную директорию и запустить парсер командой "tomitaparser.exe config.proto"(строка 11 (рис.2.9)).

Для чтения данных после первого запуска парсера (когда извлекаются атрибуты + прилагательные) был разработан метод ReadFacts(). Данный метод вычленяет слова из файла, а затем помещает их в список списков *List<List<string>> words* таким образом, чтобы данные были похожи на таблицу (см. рис.2.10):

На рис.2.10 внесены данные, которые бы были получены из текста «Сегодня прекрасная погода, но ужасное и мерзкое настроение» на основе двух этапов работы с парсером. В первый этап были бы получены известные факты (рис.2.7), во второй этап строки с прилагательными проверяются на количество слов, если нужно, разбиваются на отдельные слова. Номер цепочки определяется по существительному (атрибуту). То есть все прилагательные, относящиеся к одному существительному будут иметь один номер. Если атрибут встречается несколько раз, то такой подход позволит записать все прилагательные к одной цепочке, а значит, к одному атрибуту. Это позволит избежать ситуации, при которой атрибут, являющийся одним и тем же существительным, будет рассматриваться как несколько разных атрибутов.

Листинг2.1

StartTomita()

```
1         private void StartTomita()
2         {
3             Process cmd = new Process();
4             cmd.StartInfo.FileName = "cmd.exe";
5             cmd.StartInfo.RedirectStandardInput = true;
6             cmd.StartInfo.RedirectStandardOutput = true;
7             cmd.StartInfo.CreateNoWindow = true;
8             cmd.StartInfo.UseShellExecute = false;
9             cmd.Start();
10            cmd.StandardInput.WriteLine(@"cd C:/tomita");
11            cmd.StandardInput.WriteLine(@"tomitaparser.exe
                 config.proto");
12            cmd.StandardInput.Flush();
13            cmd.StandardInput.Close();
14            cmd.Close();
15        }
```

Рис.2.9. StartTomita()

Aa Слово	Adj/Noun	№ подцепочки	Тональность
Погода	Noun	1	
Прекрасный	Adj	1	
Настроение	Noun	2	
Ужасный	Adj	2	
Мерзкий	Adj	2	

Рис.2.10. Схематичное изображение структуры для хранения слов List<List<string> > words

2.3. Присвоение тональности атрибутам

Для удобного использования тонального словаря был разработан метод *getTonalDictionary();*, который запускается вместе с запуском приложения и переносит пары «слово» - «численное значение тональности» в *Dictionary<string, float> wordsTones*. После того, как введенный пользователем текст будет обработан парсером, каждому слову из *List<List<string> > words* будет присвоена тональность из *Dictionary<string, float> wordsTones*. После этого *List<List<string> > words* можно визуализировать как на рис.2.11:

Aa Слово	Adj/Noun	№ подцепочки	Тональность
Погода	Noun	1	0.03
Прекрасный	Adj	1	1.0
Настроение	Noun	2	0.02
Ужасный	Adj	2	-1.0
Мерзкий	Adj	2	-1.0

Рис.2.11. Схематичное изображение структуры для хранения слов List<List<string> > words

Если слово не было найдено в словаре тональности, то тогда ему присваевается значение 0.

После того как тональность из словаря была присвоена всем извлеченным парсером словам, нужно разобраться какая же тональность у атрибутов. Благодаря хранению номеров цепочек, посчитать тональность атрибута не составляет труда. Тональность каждого существительного *List<List<string> > words* суммируется с тональностью прилагательных, относящихся к той же цепочке, что и существительное (атрибут), а затем эта сумма делится на количество слов. То есть тональность атрибута вычисляется как среднее арифметическое значений его тональности и тональных слов (прилагательных). Также совершается попытка определить общею тональность текста. Если тональность отрицательная отличается от положительной меньше, чем на 10% и они не равны 0, то тогда тексту присваивается «противоречивая» тональность. Если положительная или отрицательная тональность равна нейтральной, тогда делается вывод о том, что тональность не нейтральная, а положительная/отрицательная. В противных случаях, тональность определяется по большинству атрибутов, имеющих ту или иную тональность.

2.4. Разработка графического интерфейса

Согласно сделанным ранее утверждениям в Главе 1 1.5, прототип, проводящий тональный анализ текста, должен иметь графический интерфейс. Для создания интерфейса использовались следующие компоненты windows forms: Button, Label, GroupBox, RichTextBox и DataGridView. На рис.2.12 проилюстрировано приложение с указанием на используемые компонены.

На рис.2.13 демонстрация работы созданного прототипа для анализа тональной информации с использованием Tomita-парсера и тонального словаря.

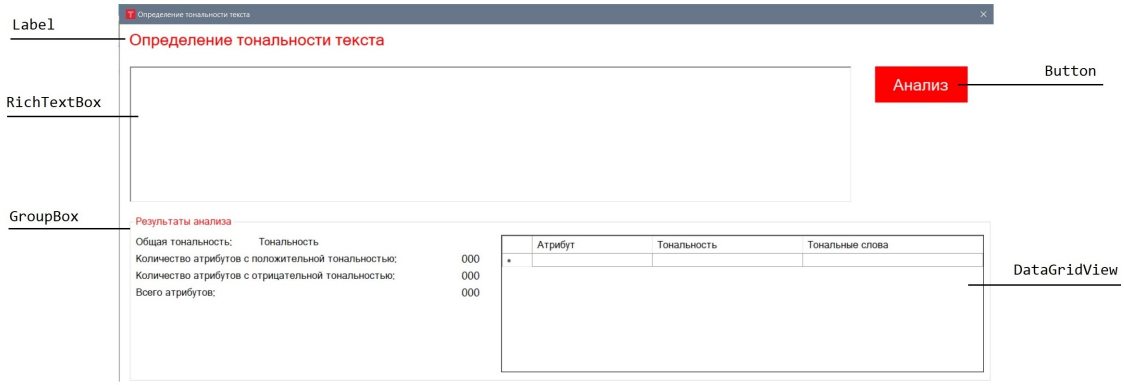

Рис.2.12. Скриншот интрефейса разработанного приложения с подписями компонентов

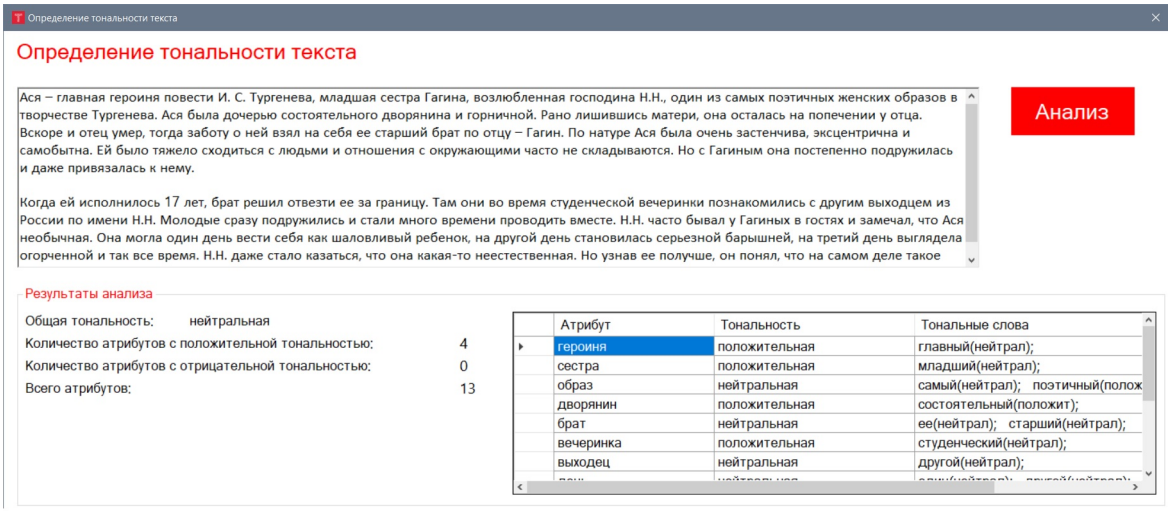

Рис.2.13. Демонстрация работы приложения

2.5. Тестирование разработанного прототипа

Тестирование разработанного прототипа будет присходить путем сравнения ожидаемых результатов и полученных. Для тестирования работы приложения выделены следующие ситуации:

- однородные прилагательные перед существительным
- неоднородные прилагательные перед существительным
- одно и то же существительное употребляется в тексте множество раз с разными прилагательными
- наличие прилагательных, подвергшихся процессу субстантивации
- отсутствие атрибутов, имеющих тональные слова
- наличие в тексте атрибутов имеющих как позитивную тональность, так и негативную (примерно в равном количестве)

Сначала было осуществлено тестирование при наличии однородных прилагательных перед существительным. Члены предложения, которые отвечают на

один и тот же вопрос и относятся к одному и тому же определяемому слову, называются однородными. [6]. Однородные члены предложения отделяются друг от друга запятой или союзом.Рассмотрим предложение: «По натуре Ася застенчивая, эксцентричная и самобытная девушка». Очевидно, что в данном предложении прилагательные «застенчивая, эксцентричная и самобытная» относятся к атрибуту «девушка». На рис.2.14 продемонстрирован результат работы приложения с этим предложением.

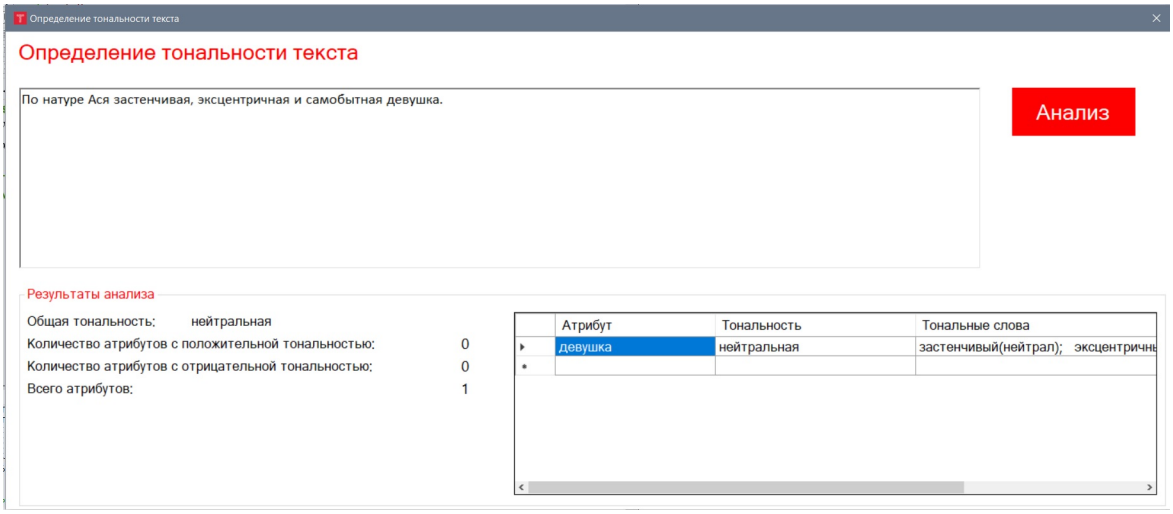

Рис.2.14. Результат работы приложения при начилии однородных прилагательных перед существительным

На рис.2.15 приведен полный список извлеченных прилагательных. Все слова, которые должны быть извлечены для атрибута «девушка» были извлечены.

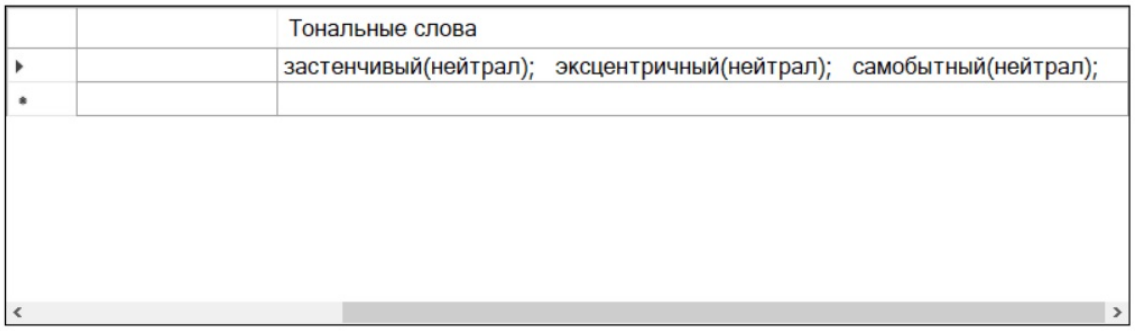

Рис.2.15. Полный список тональных слов, полученных в результате работы приложения при начилии однородных прилагательных перед существительным

Этот этап тестирования можно считать пройденным успешно.

Следущющим этапом будет произведено тестирование при наличии неоднородных прилагательных перед существительным. Неоднородные члены предложения не произносятся с перечислительной интонацией, между ними обычно нельзя поставить союз и, в отличии от однородных членов, рассмотренных в **??** [1].

Пример предложения, содержащего неоднородные прилагательные: «Ася – главная героиня повести И. С. Тургенева, один из самых поэтичных женских образов в творчестве Тургенева.». В таком предложении можно выделить атрибуты «героиня» и «образ». Неоднородные прилагательные относятся к атрибуту «образ»: «самый», «поэтичный», «женский». Прилагательные идут подряд и разделены пробелами. На рис.2.16 продемонстрирован результат работы приложения с этим предложением.

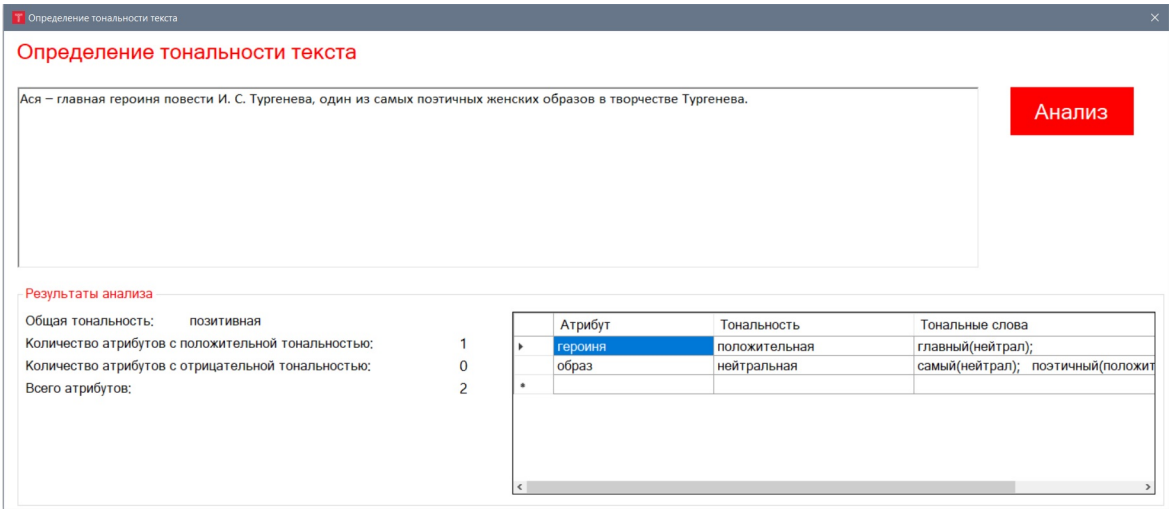

Рис.2.16. Результат работы приложения при начилии неоднородных прилагательных перед существительным

На рис.2.17 приведен полный список извлеченных прилагательных. Все слова, которые должны быть извлечены для атрибута «образ» были извлечены, на рис.2.17 они выделены синим цветом.

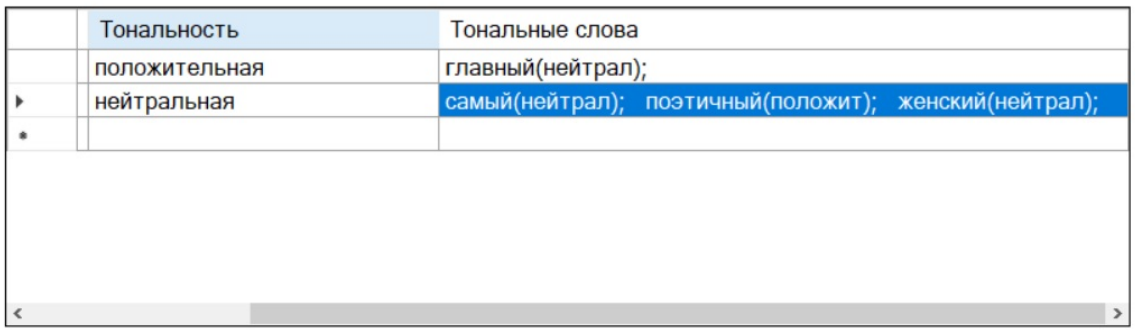

Рис.2.17. Полный список тональных слов, полученных в результате работы приложения при начилии неоднородных прилагательных перед существительным

Далее будет проведено тестирование при многократном упоминании атрибута в тексте. В данном блоке тестирования необходимо проверить действительно ли все прилагательные, относящиеся к одному и тому же существительному извлекаются? Меняется ли тональность атрибута, если какое-то тональное слово повторяется? Текст для тестирования: «Когда книга интересная, то читать ее легко

и приятно. Этого нельзя сказать о скучных книгах. Во время чтения интересной книги можно потерять счет времени.». Атрибутом, имеющим тональные слова, в тексте выступает слово «книга». Это слово употребляется как в единственном, так и множественном числе. Также к атрибуту относятся прилагательные «скучный» (один раз) и «интересный» (два раза). Очевидно, что слово «скучный» имеет негативную тональность, в то время как «интересный», наоборот, позитивную. Получается, у атрибута «книга» должна быть позитивная тональность, так как позитивных тональных слов в два раза больше, чем негативных. На рис.2.18 демонстрация результата работы приложения с упомянутым текстом.

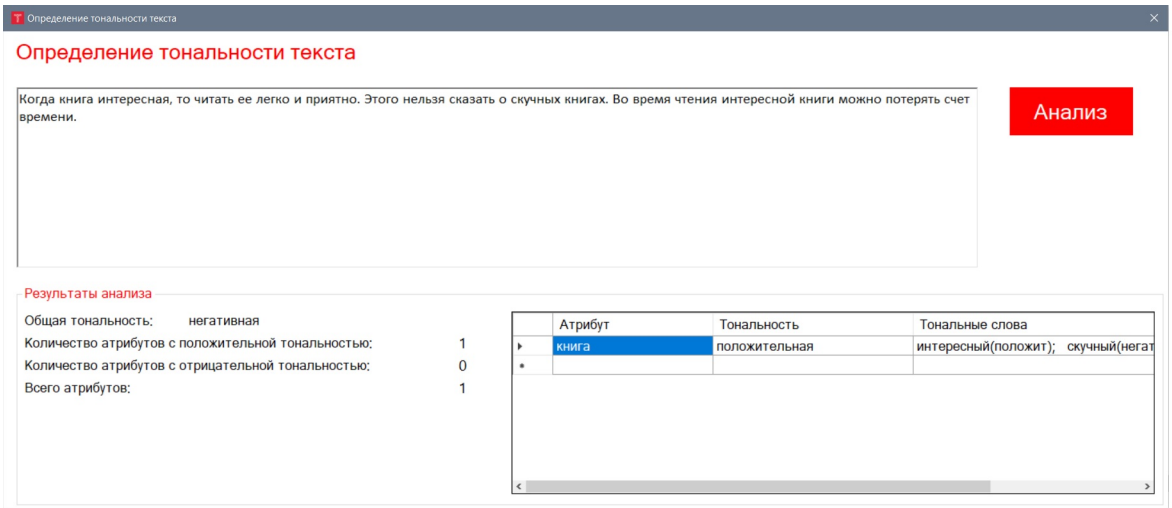

Рис.2.18. Тестирование при многократном упоминании атрибута в тексте

На рис.2.19 приведен полный список извлеченных тональных слов. Прилагательное «интересный» приведен в списке дважды и согласно рис.2.18 тональность атрибута была определена как положительная.

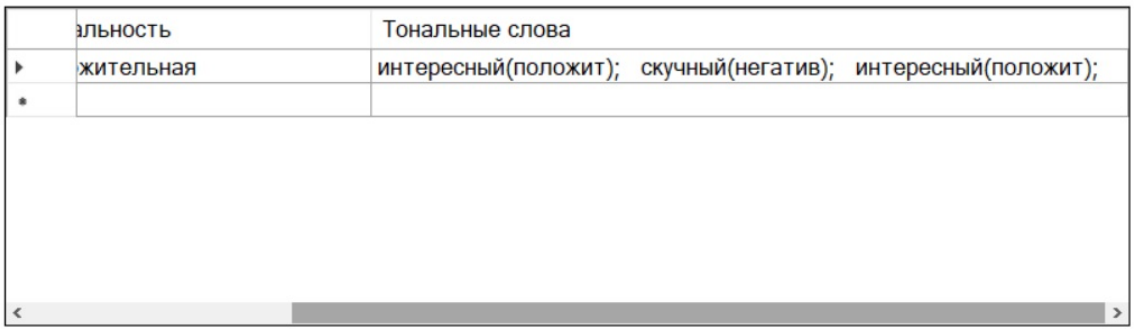

Рис.2.19. Полный список тональных слов, полученных в результате работы приложения при многократном упоминании атрибута в тексте

Следовательно, все тональные слова атрибуты были получены и учтены при расчете общей тональности атрибута.

Следующим этапом проводится тестирование при наличии прилагательных, подвергшихся процессу субстантивации. *Субстантивация* (от лат. substantivum—существительное) - это переход слов других частей речи в разряд имен существительных [9]. Например, «больной поправился» - субстантивация прилагательного «больной». Важно, чтобы при работе приложения прилагательные, которые подвержены субстантивации, не определялись как существительные, если они относятся к атрибутам. Так как процесс получения прилагательных происходит в два этапа: сначала получение парсером из начального текста цепочек существительное + прилагательные, к нему относящиеся, а потом на вход парсеру подается строка с извлеченными прилагательными, то важно, чтобы прилагательные во время второго запуска парсера оставались прилагательными для парсера. Теститирование будет проводиться на следующем тексте: «Уставший рабочий сел отдыхать на лавочку. Он отряхнул свой рабочий комбинезон и посмотрел вдаль.». В этом тексте слово «рабочий» выступает в роли как существительного, так и прилагательного. Но очевидно, что атрибутами, имеющими тональные слова, являются слова «рабочий» и «комбинезон». Демонстрация результата запуска приложения с вышеупомянутым текстом представлена на рис.2.20.

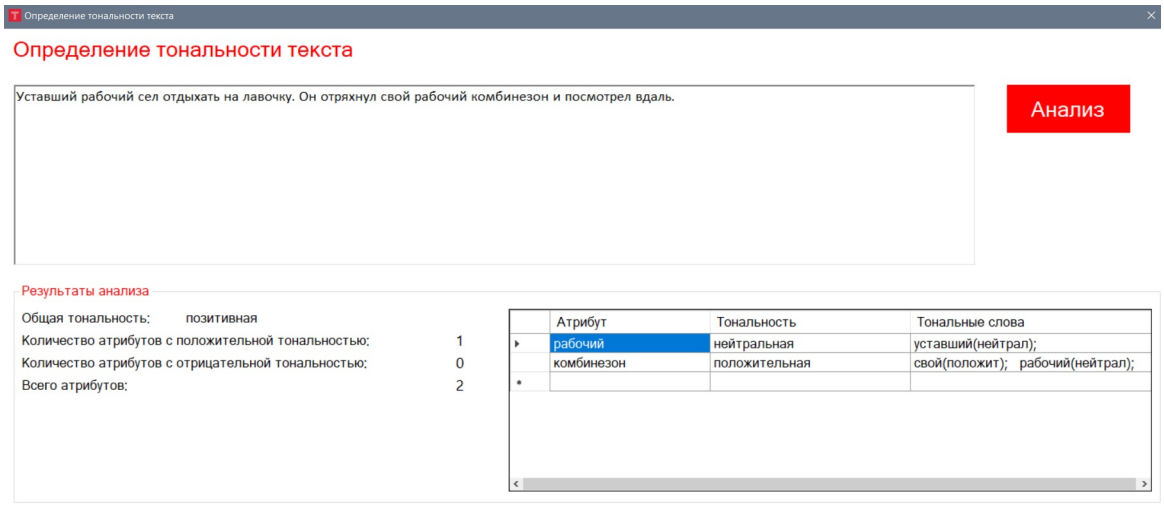

Рис.2.20. Тестирование при наличии прилагательных, подвергшихся процессу субстантивации

Приложение корректно обработало текст и правильно к атрибуту существительное,а прилагательное, как и ожидалось, было записано в тональные слова.

Далее проведено тестирование при отсутствии атрибутов, имеющих тональные слова. При отсутствии в тексте атрибутов, имеющих тональные слова, необходимо сообщить пользователю об этом. Демонстрация работы программы при вводе текста «Сегодня цветок завял», не содержащего в себе прилагательных, представлена на рис.2.21.

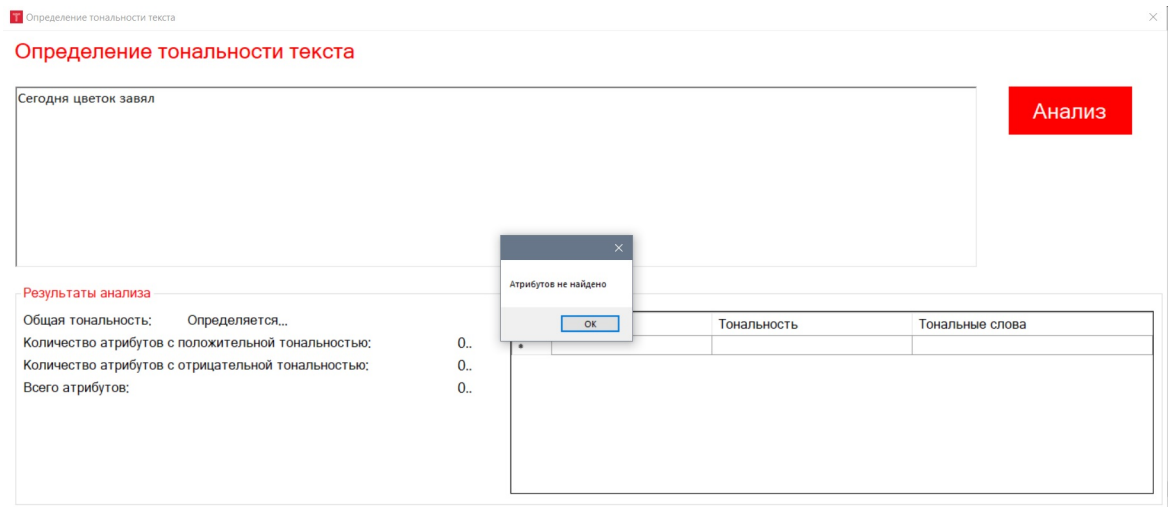

Рис.2.21. Тестирование при отсутствии атрибутов в тексте

Как видно на рис.2.21, после нажатия на кнопку «Анализ» появляется Message Box, уведомляющий о том, что атрибутов не найдено. После нажатия на «ок» тексту присваивается нейтральная тональность - см. рис.2.22

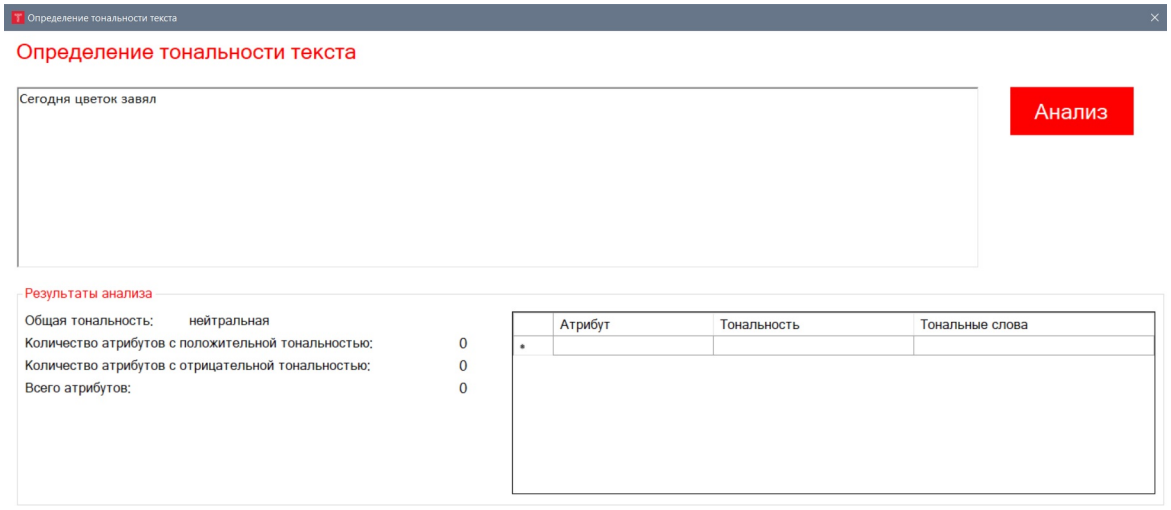

Рис.2.22. Определение тональности при отсутствии атрибутов в тексте

Следовательно, при отсутствии атрибутов, имеющих тональные слова, приложение уведомляет пользователя и присваивает тексту нейтральную тональность, - это являлось желательным результатом.

Далее проводится тестирование при наличии атрибутов, имеющих противоположную тональность. Тестирование будет проводиться при использования текста, упоминавшйся ранее при исследовании Томита-парсера 2.2: «Сегодня прекрасная погода, но ужасное настроение». Тональность данного предложения затруднительно определить однозначно. Поэтому желаемым результатом является корректная оценка тональности каждого из атрибутов («погода» - позитивная

тональность, «настроение» - негативная) и вывод о противоречивой тональности всего текста. Результат тестирования можно наблюдать на рис.2.23.

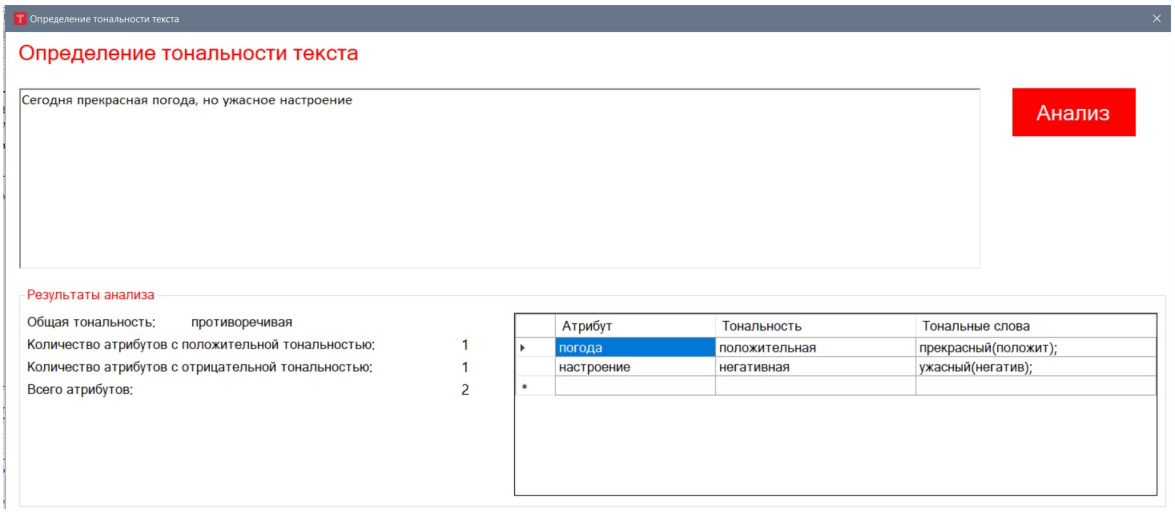

Рис.2.23. Тестирование при отсутствии атрибутов в тексте

Как видно на рис.2.23 полученные результаты полностью совпали с желаемыми.

Согласно тестированию, которое заключалось в сопоставлении желаемых результатов и получаемых, приложение успешно справляется с задачей определения тональности атрибутов и имеет гибкую систему оценивания, что позволяет максимально точно оценить тональность текста. Следовательно, результаты тестирования удовлетворительные по каждому из рассматриваемых пунктов.

ЗАКЛЮЧЕНИЕ

Анализ тональности имеет важное практическое применение и поиск решений для его реализации имеет большой интерес от представителей самых разных отраслей. Существующие на данный момент методы оценки тональности текста имеют ряд недостатков, поэтому попытка разработать свой метод была обоснована. Использованный в работе Томита-парсер имеет большую функциональность. Благодаря существующей документации [11], которая включает в себя как текстовые объяснения, так и видеоролики, разработка правил для работы с парсером не представляет существенных сложностей. Для достижения поставленной цели, а именно - разработка прототипа с применением Tomita-парсера для анализа атрибутивной тональности текстовых сообщений были пройдены следующие этапы:

- рассмотрены методы оценки тональности, выявлены их достоинства и недостатки
- разработан метод для оценки тональности на основе рассмотренных методов
- разработан прототип в виде оконного приложения

Цель была достигнута, что подтверждается успешно проведенным тестированием разработанного прототипа. Для улучшения разработанного приложения можно разработать правила для Томиты-парсера, позволяющего извлекать слова других частей речи (например, наречия), чтобы повысить полноту оценивания тональности текста.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

[1] *2.3.2. Однородные и неоднородные определения*. URL: https://licey.net/free/4-russkii_yazyk/40-kurs_russkogo_yazyka_sintaksis_i_punktuaciya/stages/713-232_odnorodnye_i_neodnorodnye_opredeleniya.html (дата обращения: 24.06.2021).

[2] *Анализ тональности текста*. URL: https://clck.ru/V9Dn9 (дата обращения: 06.06.2021).

[3] *Анализ тональности текста с использованием методов машинного обучения*. URL: http://ceur-ws.org/Vol-2233/Paper_8.pdf (дата обращения: 06.06.2021).

[4] *Извлечение объектов и фактов из текстов в Яндексе. Лекция для Малого ШАДа*. URL: https://habr.com/ru/company/yandex/blog/205198/ (дата обращения: 06.06.2021).

[5] *Классификация текстов и анализ тональности*. URL: https://clck.ru/SNgQv (дата обращения: 06.06.2021).

[6] *Когда между прилагательными ставится запятая?* URL: https://pishugramotno.ru/punktuacia/kogda-mezhdu-prilagatelnymi-stavitsya-zapyataya (дата обращения: 24.06.2021).

[7] *Обучаем компьютер чувствам*. URL: https://habr.com/ru/post/149605/ (дата обращения: 26.06.2021).

[8] *Системная социология: Opinion Mining*. URL: https://www.isras.ru/index.php?page_id=1024 (дата обращения: 06.06.2021).

[9] *Словарь лингвистических терминов. Субстантивация*. URL: https://classes.ru/grammar/114.Rosental/17-s-3/html/unnamed_68.html (дата обращения: 24.06.2021).

[10] *Томита-парсер*. URL: https://yandex.ru/dev/tomita/ (дата обращения: 11.06.2021).

[11] *Томита-парсер учебник*. URL: https://yandex.ru/dev/tomita/doc/tutorial/concept/about.html (дата обращения: 06.06.2021).