

САНКТ-ПЕТЕРБУРГСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

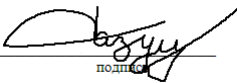
Кафедра технологии программирования

Курсовая работа на тему:

**АСПЕКТНО-ОРИЕНТИРОВАННЫЙ СЕНТИМЕНТ АНАЛИЗ ПОЛЬЗОВАТЕЛЬСКИХ СООБЩЕНИЙ  
В СОЦИАЛЬНЫХ СЕТЯХ**

Студент:

Разумилов Егор Сергеевич,  
студент группы 18.Б08-пу

  
\_\_\_\_\_

Научный руководитель:

Блеканов Иван Станиславович,  
кандидат технических наук, доцент

  
\_\_\_\_\_

Санкт-Петербург  
2021

# Содержание

<b>1</b>	<b>Введение</b>	<b>3</b>
<b>2</b>	<b>Классификация существующих методов</b>	<b>4</b>
<b>3</b>	<b>Извлечение аспектных терминов</b>	<b>5</b>
3.1	Частотные методы . . . . .	5
3.2	Синтаксический подход . . . . .	5
3.3	Машинное обучение с учителем . . . . .	6
3.4	Машинное обучение без учителя . . . . .	6
3.5	Сравнение методов . . . . .	7
<b>4</b>	<b>Классификация полярности аспектов</b>	<b>7</b>
4.1	Словарные методы . . . . .	7
4.2	Машинное обучение с учителем . . . . .	7
4.3	Машинное обучение без учителя . . . . .	10
4.4	Сравнение методов . . . . .	10
<b>5</b>	<b>Методы, решающие и АТЕ, и АРС</b>	<b>10</b>
5.1	Синтаксические методы . . . . .	10
5.2	Машинное обучение с учителем . . . . .	10
5.3	Машинное обучение без учителя . . . . .	13
5.4	Сравнение методов . . . . .	13
<b>6</b>	<b>Результаты работы</b>	<b>14</b>

# 1 Введение

Задача сентимент анализа пользовательских сообщений, то есть определение эмоциональной оценки мнений, высказанных пользователями, является актуальной для множества сфер: маркетинг (в социальных сетях активно обсуждаются различные товары), социология, политология (достаточно часто необходимо знать отношение людей к тому или иному событию или личности) и так далее.

Задача сентимент анализа, как присвоение целому высказыванию пользователя метки «положительное», «нейтральное» и «негативное» является задачей широко изученной. Но достаточно очевиден тот факт, что пользователь в своем сообщении может говорить как и в негативном, так и в позитивном ключе, например, отзыв «Этот телефон прекрасен, за исключением его отвратительной камеры» не может быть однозначно промаркировано. Таким образом, задачу расширяют до *аспектного сентимент анализа*:

**Для пользовательского сообщения найти все упорядоченные пары типа  $(s, a)$ , где  $s \in \{-1, 0, 1\}$  — сентимент, число «1» соответствует положительному описанию аспекта, «0» - нейтральному, «-1» - негативному,  $a$  - слово или фраза из пользовательского сообщения, которое отображает объект, подвергающийся оценке (будем называть это *аспект*)**[1].

Данная постановка задачи позволяет достаточно полно описывать эмоциональную оценку пользователя. И такая точная характеристика позволит, например, маркетологам, понимать отрицательные стороны продукта компании для дальнейшего его улучшения.

Таким образом, **целью** данной работы является обзор существующих методов аспектного сентимент анализа и проведение сравнительного анализа.

## 2 Классификация существующих методов

Задачу аспектного сентимент анализа можно разделить на две подзадачи (в соответствии с [2]):

- **Извлечение аспектных терминов** (Aspect Term Extraction, ATE) - задача из раздела распознавания именованных сущностей. Необходимо определить в пользовательском сообщении, что является аспектом. Если представлять эту задачу с помощью IOB разметки, то, например, в сообщении «The price is reasonable although the service is poor» правильной IOB разметкой будет  $\{O, B_{asp}, O, O, O, O, B_{asp}, O, O\}$ .
- **Классификация полярности аспектов** (Aspect Polarity Classification, APC) - каждому аспекту пользовательского сообщения присваивается некоторая эмоциональная окраска.

Многие существующие работы решают только одну из этих подзадач, на основании [1]. Поэтому сначала мы поговорим о решении задачи ATE. Существует несколько разных подходов к ее решению:

- **Частотные** (frequency-based) - словами-аспектами признаются слова, которые часто встречаются в тексте (обычно обрабатываются существительные, простые или составные)
- **Синтаксические** (syntax-based) - для предложений строятся синтаксические деревья, которые далее подвергаются разбору по набору некоторых правил.
- **Обучение с учителем** (supervised learning).
- **Обучение без учителя** (unsupervised learning).

Задача APC также решается с помощью обучения с учителем и без него, и вместо частотных и синтаксических подходов используются **словарные методы** (dictionary based).

Также существуют методы, которые решают обе эти задачи вместе. Такие методы также представлены синтаксическими подходами и машинным обучением.

Далее будут рассмотрены каждый подходы в отдельности.

## 3 Извлечение аспектных терминов

### 3.1 Частотные методы

Одной из самых известных работ в этой области является исследование [3]. В данной работе представляется многоступенчатая система по частотному выделению аспектов. Для начала происходит *POS-тэггинг* (у каждого слова определяется его часть речи). Далее, находятся самые частые простые или составные существительные, используя алгоритм, предложенный в статье [4]. Последний применялся для поиска так называемых *ассоциативных правил*, которые можно определить как:

Пусть  $I = \{i_1, \dots, i_m\}$  - какое-либо множество, а  $D$  - набор транзакций (датасет). Тогда ассоциативное правило можно определить как отображение  $X \rightarrow Y$ , где  $X \subset I, Y \subset I, X \cap Y = \emptyset$ .

В реализацию поиска таких правил включался алгоритм нахождения достаточно часто появляющихся подмножеств длины  $k$ . В работе [3] использовалось  $k = 1, \dots, 3$ , так как делается допущение, что аспект не будет длины больше 3. Далее, авторы метода убирают излишние фразы, которые могли попасть в набор аспектов по ошибке, с помощью 2 методов:

- **Отсечение по компактности** (Compactness pruning) - если не нашлось по крайней мере 2 отзыва, в которых слова из анализируемой фразы встречаются достаточно близко, то эта фраза убирается из итогового набора аспектов.
- **Отсечение по избыточности** (Redundancy pruning) - если аспект слишком часто встречается в высказываниях в составе другой фразы, то есть смысл рассматриваемый аспект убрать из итогового набора аспектов.

Далее происходит поиск эмоционально окрашенных слов, применяющихся к аспекту. В рассматриваемой работе приводится достаточно простой метод поиска: ищутся ближайшие к аспекту *прилагательные*. После этого такие слова используются для поиска нечастых аспектов. Делается предположение, что пользователи пользуются одинаковыми эмоционально окрашенными словами, чтобы описать как часто, так и редко обсуждаемые аспекты. На основе этой гипотезы поиск проводится следующим образом: если в высказывании не нашлось достаточно частых аспектов, но присутствуют эмоционально окрашенные слова, то ищутся существительные, ближайшие к этим словам. Они и будут аспектами в этом высказывании.

Одним из недостатков этого метода является то, что он выделяет только явные аспекты. Например, в предложении «While light, it will not easily fit in pocket» явно говорится о недостатках *размера* товара, но явно этого в высказывании не указано. В работе [5] решается эта проблема тем же поиском ассоциативных правил, только на них накладываются условия того, что множество  $X$  в определении - это эмоциональные слова, а  $Y$  - аспекты. И такие правила ищутся не из изначального датасета высказываний пользователей, а из матрицы совместных встречаемостей «слово-сентимент: явный аспект». Таким образом, для получения неявных аспектов, работа [5] предполагает наличие явных аспектов. Поэтому два этих метода могут работать в паре.

Еще одним явным недостатком частотного подхода является то, что часто встречающиеся существительные в высказываниях в действительности могут не быть аспектами. Эту проблему решает работа [6], которая сравнивает частоту появления существительного в высказываниях в датасете с так называемой базовой частотой, основанной на дополнительно собранном датасете из 100 миллионов английских слов. Если частота существительного больше базовой, то последнее рассматривается как аспект.

### 3.2 Синтаксический подход

Метод основан на построении синтаксического дерева высказывания, с последующим его анализом по некоторым правилам. Например, достаточно простой паттерн - это встречающиеся рядом эмоционально окрашенное прилагательное и существительное, например, в словосочетании «fantastic food». В таком случае слово «food» очевидно относится к аспектам. Плюсом этого метода является возможность достаточно просто находить редко упоминаемые аспекты. Но для хорошей работы такого алгоритма необходимо учитывать достаточно большое число различных правил языка, что может быть проблематично. В исследовании [7] применяются 2 метода улучшения стандартного rule-based алгоритма:

- Применяются 2 эвристики, которые позволяют собрать некоторые частные части речи в предложении в одну структуру, и общая логика синтаксического дерева не поменяется (например, 2 части речи NP, одна из которых является предком, а другая - непосредственным потомком, объединяются в одну часть речи NP). Или, например, возможна замена слова с частью речи NNS на просто NN.

- Происходит построение шаблонов синтаксических деревьев, по которым можно вычислить словосочетание-аспект. На неразмеченных данных также строится синтаксическое дерево, и выделяются некоторые его поддеревья (у шаблонов на размеченных данных также происходит разбиение на поддеревья). После этого, количество принципиально разных поддеревьев подсчитывается и заносится в вектор-подобную структуру вида:  $\Phi(T) = (\phi_1(T), \dots, \phi_n(T))$ , где функция  $\phi_i(T)$  дает количество поддеревьев типа  $i$  в дереве  $T$ . Далее, такие вектора сравниваются с векторами для уже готовых шаблонов. Если пройден порог по схожести, то словосочетание, находящееся в этом шаблоне, признается аспектом.

В работе [8] представлен несколько иной метод выделения аспектов по синтаксическим деревьям. В нем сначала вручную выделяется несколько начальных слов, выражающих эмоции. Далее, строится несколько шаблонных синтаксических деревьев, которые связывают какую-либо из пар:

- Сентимент-слово и аспект;
- Сентимент-слово и сентимент-слово;
- Аспект и аспект;

Таким образом, с помощью изначально выделенных сентимент-слов и синтаксических правил, заданных шаблонными деревьями, выделяются дополнительные сентимент-слова и аспекты. Процесс продолжается до тех пор, пока набор сентимент-слов и аспектов перестанет пополняться. Плюсом этого метода является то, что ему необходимо лишь малое число начальных эмоционально окрашенных слов для правильно работы.

### 3.3 Машинное обучение с учителем

До серьезного развития глубоких нейронных сетей подходов решения задачи АТЕ, основанных на машинном обучении с учителем, было не очень много. Одна из таких, например, работа [9] использует условные случайные поля (CRF) на нескольких признаках, среди которых само слово, его часть речи, связано ли это слово с каким-либо сентимент-словом, и несет ли вообще анализируемое предложение какой-либо сентимент. В статье авторов показаны метрики их метода, которые гораздо меньше, чем даже у простых частотных методов.

В 2018-2019 годах, когда широкое распространение получили глубокие нейронные сети, такие как BERT, было совершено большое число исследований, которые приводят модели машинного обучения с учителем, решающие задачу аспектного сентимент анализа с достаточно хорошей точностью. Но эти методы решают либо задачу APC, либо полную задачу сентимент анализа, поэтому рассмотрены они будут позднее.

### 3.4 Машинное обучение без учителя

Одним из самых популярных методов, применяемых к задаче АТЕ, является *латентное размещение Дирихле* (LDA). Этот алгоритм в чистом виде применяется для тематического моделирования. Делается предположение, что темой высказывания в некотором роде аспект и является, делаются попытки применить LDA и к задаче извлечения последних. Применение чистого LDA не на документе, а на отдельных предложениях затруднительно, так как в таком случае «мешок слов», на котором и основана работа LDA, будет слишком маленьким и приемлемого решения не получится.

В работе [10] предлагается новая версия алгоритма LDA - *Multi-grained LDA*. Данный алгоритм представляет текст как набор скользящих окон, включающих в себя сразу несколько предложений. Таким образом, алгоритм позволяет получать сразу 2 типа тем: глобальные и локальные. Локальные темы и являются аспектами.

В статье [11] применяется LDA алгоритм вкупе со скрытыми марковскими полями для выделения слов-аспектов. Такое выделение возможно с помощью применяемых синтаксических зависимостей, которые выделяют аспектные слова.

Понятно, что методы, основанные на LDA, годятся только для больших текстов, короткое пользовательское сообщение, очевидно, даст куда более плохой результат.

### 3.5 Сравнение методов

Сравнительная таблица по разобранным методам решения задачи АТЕ представлена ниже. В каждой статье использовались свои датасеты для оценки качества, так что последняя графа по большей части показывает своеобразный «максимум», который может выдать модель.

Ссылка на метод	Класс метода	Метрики качества
[3]	Frequency-based	precision: 72%, recall: 80%
[5]	Frequency-based	precision: 76.29%, recall: 72.71% , $F_1$ : 74.46%
[6]	Frequency-based	precision: 85%-90%
[7]	Syntax-based	precision: max. 76% recall: max. 68%
[8]	Syntax-based	precision: 88%, recall: 83%
[9]	Supervised machine learning	precision: 74.9%, recall: 66.1%
[10]	Unsupervised machine learning	— (авторы просто показывают примеры работы)
[11]	Unsupervised machine learning	precision: 83.33%, recall: 81.12%

## 4 Классификация полярности аспектов

Эта подзадача покрывается гораздо большим числом исследований. Ниже будут приведены некоторые достаточно известные и значимые работы в этой области, подразделенные на 3 класса.

### 4.1 Словарные методы

На самом деле, методы из этого класса пользуются одной идеей: модель имеет доступ к таким ресурсам, как, например, WordNet, который позволяет искать синонимы к сентимент-словам. Например, в работе [12] также используется словарь, который ранжирует прилагательные по сентименту. Неизвестные прилагательные ищутся в WordNet в графе синонимов, поиском в ширину находят ближайшие, и по методу ближайших соседей и неизвестному прилагательному присваивается какой-либо сентимент.

### 4.2 Машинное обучение с учителем

До скачка в развитии нейронных сетей для задачи сентимент анализа использовались стандартные и всем известные подходы машинного обучения как, например, *Support Vector Machine* (SVM). В этом плане примечательным исследованием является работа [13]. В статье приводится не столько сентимент анализ заранее выделенных аспектов, сколько анализ коротких выражений, что также может быть полезно в нашем исследовании. Также важно то, что SVM в этом исследовании получает не просто «мешок слов», а несколько признаков, таких как:

- Само слово, а также его лемму;
- Данные по сентименту слова: является ли оно эмоциональным, является ли оно отрицающим, передается его полярность в соответствии с несколькими словарями;
- И также передается значение специальной эвристики, как голосование полярностей: считается число слов с положительным и отрицательным оттенком, и далее, каких больше, таким и будет оттенок высказывания (также учитывается число слов-отрицателей, которые, скорее всего, «переворачивают» сентимент высказывания).

Также авторы разработали еще одну эвристику, так называемую композитную семантику: вручную был определен набор правил для последовательного объединения сентимент слов с получением итогового сентимента. Эти правила были встроены в алгоритм обновления SVM.

Одни из первых попыток использования глубокого обучения в задаче аспектного сентимент-анализа — это применение рекуррентных нейронных сетей, как, например, в работе [14]. Формально авторы определили задачу так:

Имеется пара  $(\mathbf{w}^\tau, \mathbf{w})$ , где  $\mathbf{w}^\tau = \{w_1^\tau, \dots, w_m^\tau\}$  - подпоследовательность  $\mathbf{w} = \{w_1, \dots, w_n\}$  - последовательности слов в высказывании, у каждого слова определен эмбединг (представление слова в векторном пространстве)  $\mathbf{x}^\tau = \{x_1^\tau, \dots, x_m^\tau\}$ ,  $\mathbf{x} = \{x_1, \dots, x_n\}$ . Необходимо относительно аспекта  $\mathbf{w}^\tau$  оценить полярность высказывания  $\mathbf{w}$ .

Авторы статьи на основе нейронной сети LSTM разработали свою сеть под названием TNet, принцип которой изображен на рисунке 1.

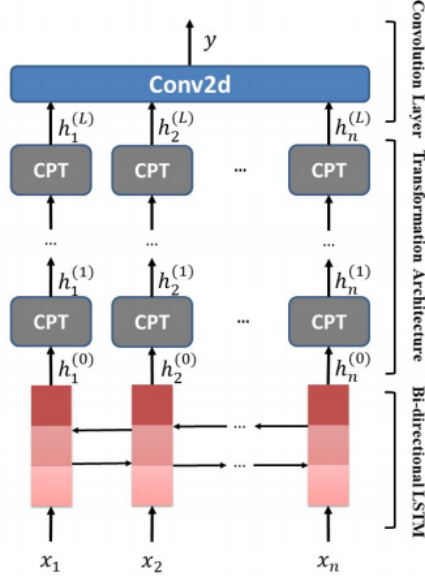


Рис. 1: Устройство TNet.

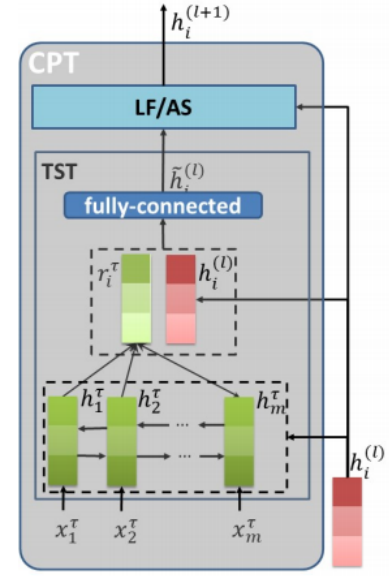


Рис. 2: Устройство модуля CPT.

Сначала из эмбедингов всего текста получаются вектора  $h_i^{(0)}, i \in [1, n]$ , так называемые контекстуализированные эмбединги слов, с помощью двунаправленной сети LSTM. Далее, эти эмбединги подаются в модуль CPT (Context-Preserving Transformation, трансформация с сохранением контекста, его устройство показано на рисунке 2), который для начала применяет другую двунаправленную LSTM только на слова из аспекта, получая контекстуализированные эмбединги  $h_i^\tau, i \in [1, m]$  и далее каждое такое представление ассоциируется с вектором  $h_i^{(l)}$ , который и был подан в блок CPT, на основе этих ассоциаций строится вектор:  $r_i^\tau = \sum_{j=1}^m h_i^\tau * \mathcal{F}(h_i^{(l)}, h_j^\tau)$ , где функция  $\mathcal{F}$  отображает схожесть двух эмбедингов и вычисляется по формуле  $\mathcal{F}(h_i^{(l)}, h_j^\tau) = \frac{\exp(h_i^{(l)\top} h_j^\tau)}{\sum_{k=1}^m \exp(h_i^{(l)\top} h_k^\tau)}$ . Наконец, вектора  $r_i^\tau$  и  $h_i^{(l)}$  конкатенируются и подаются в полносвязный слой с нелинейной функцией активации и получается вектор  $\tilde{h}_i^{(l)}$ . Далее необходимо провести преобразование над получившимся вектором, чтобы сохранить контекст. Для этого авторы разработали 2 схемы:

- **Пересылка без потерь** (Lossless Forwarding, LF): вектор выхода из CPT  $h_i^{(l+1)}$  получается с помощью простой суммы:  $h_i^{(l+1)} = \tilde{h}_i^{(l)} + h_i^{(l)}, i \in [1, n], l \in [0, L]$ .
- **Адаптивное масштабирование** (Adaptive Scaling, AS): такой способ позволяет веса входа и признаков изменять динамически, для этого определяется так называемая gating function (можно перевести как «потококонтролирующая функция»), вводится переменная  $t_i^{(l)} = \sigma(W_{trans} h_i^{(l)} + b_{trans})$ , где  $\sigma$  - сигмоидная функция активации. Таким образом, итоговая формула для выхода из модуля CPT:  $h_i^{(l+1)} = t_i^{(l)} \odot \tilde{h}_i^{(l)} + (1 - t_i^{(l)}) \odot h_i^{(l)}$ , где символ  $\odot$  обозначает поэлементное умножение.

Одна из этих двух схем применяется на выходе из CPT, и результат работы снова подается в тот же модуль. Далее, когда каждый эмбединг проходит L раз CPT, то все вектора выходов подаются в сверточную нейросеть, на основе которой и производится классификация по сентименту.

С появлением таких модулей машинного обучения, как attention и self-attention, и на основе этих модулей таких предобученных нейросетей, как BERT, архитектуры для решения задачи APC начали основываться на них, что, конечно же, повысило и метрики качества.

Например, в работе [15] было разработано сразу несколько моделей: было предложено модифицировать ввод в стандартную модель BERT (как известно, BERT обучалась определять, является ли одно предложение логическим следованием другого, для этого в модель подавалась комбинация «[CLS] + sentence 1 + [SEP] + sentence 2 + [SEP]»), то есть подавать комбинацию токенов «[CLS] + sentence + [SEP] + aspect



+ [SEP]». Уже такое небольшое улучшение показывало state-of-the-art результаты.

Но авторы пошли дальше и разработали еще одну архитектуру, которая была названа как Attentional Encoder Network (AEN). Общая схема работы сети представлена на рисунке 3.

Сначала необходимо получить эмбединги слов предложения и аспекта. Для этого авторы рассматривают

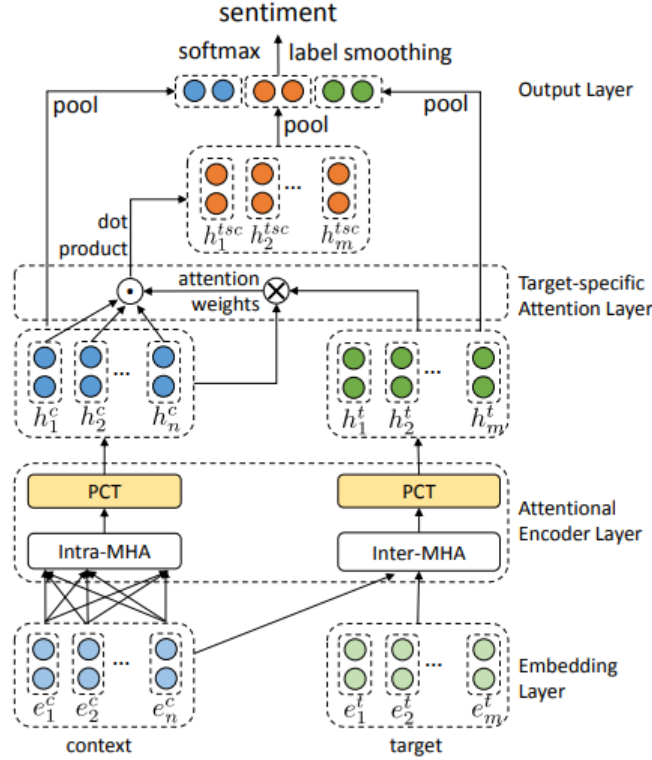


Рис. 3: Устройство сети AEN.

2 возможности: взять предобученные вектора либо из GloVe, либо из самой BERT. Впоследствии 2 эти разные архитектуры будут называться AEN-GloVe и AEN-BERT.

Далее к этим эмбедингам применяется механизм *Multi-Head Attention* (MHA). Необходимо пояснить, что такое механизм внимания:

Функция внимания отображает последовательность *ключей*  $\mathbf{k} = \{k_1, \dots, k_n\}$  и запрос  $\mathbf{q} = \{q_1, \dots, q_m\}$  в последовательность выходов  $\mathbf{o}$ :  $Attention(\mathbf{k}, \mathbf{q}) = softmax(f_s(\mathbf{k}, \mathbf{q})) \mathbf{k}$ , где функция  $f_s$  выдает семантическую схожесть между векторами  $q_j$  и  $k_i$ :  $f_s(k_i, q_j) = \tanh([k_i; q_j] \cdot W_{att})$  («;» значит конкатенация векторов). Смысл этой процедуры в том, чтобы оценить «важность» каждого слова в предложении (в данном случае ключей) относительно запроса.

Понятно, что в таком виде у механизма внимания есть недостаток: будет учитываться только один семантический смысл сравниваемых слов, что приведет к возможному упущению некоторых смыслов в тексте. Тогда была разработана архитектура Multi-Head Attention: матрицы  $\mathbf{k}$  и  $\mathbf{q}$  переводятся линейными преобразованиями в матрицы меньших размерностей, из этих матриц получаются выходы  $\mathbf{o}$ , они конкатенируются и дополнительно преобразуются линейным преобразованием:

$$MHA(\mathbf{k}, \mathbf{q}) = (\mathbf{o}^1; \dots; \mathbf{o}^{n_{head}}) \cdot W_{mh}, \mathbf{o}^h = Attention^h(\mathbf{k}, \mathbf{q}).$$

На рисунке 3 видно, что в модели используется 2 разных архитектуры MHA:

- Intra-MHA:  $\mathbf{c}^{intra} = MHA(\mathbf{e}^c, \mathbf{e}^c)$ . Анализирует вектора высказывания  $\mathbf{e}^c$  относительно себя же.
- Inter-MHA:  $\mathbf{c}^{inter} = MHA(\mathbf{e}^t, \mathbf{e}^c)$ . Анализирует вектора аспекта  $\mathbf{e}^t$  относительно векторов высказывания.

Далее выходы из обоих механизмов внимая подаются в модуль Point-wise Convolution Transformation (PCT). Формально, это функция  $PCT(\mathbf{h}) = \sigma(\mathbf{h} * W_{pc}^1 + b_{pc}^1) * W_{pc}^2 + b_{pc}^2$ , где  $\sigma$  - экспоненциальная функция активации.

После этого к двум выходам из модулей РСТ снова применяется модуль внимания, и с помощью пулинга выходы с нескольких слоев объединяются и на основе их с помощью функций *softmax* считается сентимент.

Существует еще множество различных архитектур нейронных сетей для решения задачи APC, в работе приводится только несколько из них для понимания общей схемы построения таких архитектур.

### 4.3 Машинное обучение без учителя

Одна из работ на эту тему - это система, разработанная на основе статьи [16]. Каждый аспект используется для поиска слов (фраз), выражающих сентимент, путем их поиска окрестности этого сентимента, где близость измеряется с использованием синтаксических зависимостей. Затем исследуется каждая потенциальная фраза сентимента, и сохраняются только те, которые показывают положительный или отрицательный сентимент. Конечным результатом является набор фраз сентиментов с их наиболее вероятной меткой полярности, положительной или отрицательной.

### 4.4 Сравнение методов

Нужно отметить, что словарные методы решения задачи APC требуют большого числа метаинформации, например, доступ в WordNet.

Также нужно заметить факт того, что модель SVM до изобретения нейронных сетей использовалась в обработке текстов практически повсеместно вследствие своей хорошей теоретической обоснованности. После появления нейронных сетей и в особенности BERT от этой методики немного отошли. Методы обучения без учителя используют достаточно большое число эвристик, и таким методам нужно много данных для хорошей работы.

Сравнительная таблица по метрикам качества описанных выше методов представлена ниже. Нужно напомнить, что в большинстве случаев авторы составляли свои датасеты, и поэтому метрики качества представлены, чтобы оценить возможности алгоритма.

Ссылка на метод	Класс метода	Метрики качества
[3]	Dictionary-based	accuracy: 80%
[12]	Dictionary-based	Ranking Loss: 0.49
[13]	Supervised machine learning	accuracy: 90.70%
[14]	Supervised machine learning	accuracy: max 80.79%
[15]	Supervised machine learning	accuracy: max 83.12%
[16]	Unsupervised machine learning	precision: 84.8%, recall: 89.28%

## 5 Методы, решающие и АТЕ, и APC

### 5.1 Синтаксические методы

Большинство алгоритмов совместного решения задач АТЕ и APC, основанных на синтаксическом анализе, имеют те же недостатки, что и при решении отдельно задачи АТЕ: необходимо грамотно выбирать синтаксические правила, по которым будут искаться аспекты, ведь если выбрать не слишком универсальные правила, то это приведет к высокому precision, но к низкому recall. Слишком сильная синтаксическая проработка языка повысит шанс того, что модель будет выдавать не аспекты, что приводит к низкому precision, но к высокому recall.

Одна из самых известных работ в этой области - это исследование [17]. В этой работе авторы справедливо предположили, что искать сентимент гораздо проще, чем аспект, поэтому после POS-тэггинга ищутся слова (фразы) сентимента, после этого производится синтаксический анализ текста, и на основе уже найденных слов-сентиментов, по некоторым синтаксическим шаблонам, определенным авторами вручную, производится поиск уже аспектов. Синтаксические правила, определенные авторами, были достаточно обширными, и поэтому они получили метод с очень высоким precision, но с низким recall.

### 5.2 Машинное обучение с учителем

В работе [18] подходят к задаче аспектного сентимент анализа как к задаче распознавания именованных сущностей, причем одной моделью проводится поиск и слов-аспектов, и слов-сентиментов. Например, результатом работы модели на предложении «The camera comes with a pitiful 32mb compact flash card.»

будет набор меток  $\{O, F_B, O, O, O, C_B, F_B, F_I, F_I, F_I, O\}$ , где  $F_B$  - feature beginning,  $F_I$  - feature inside,  $C_B$  - negative beginning,  $O$  - other.

Авторы применяют CRF для составления модели классификации, причем структура CRF была выбрана как на рисунке 4: Такая организация условного случайного поля обусловлена тем, что дерево хорошо

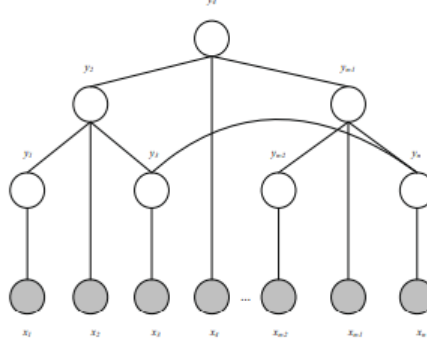


Рис. 4: Структура CRF.

описывает синтаксические зависимости в предложениях, а связи между вершинами с разными предками отражает важность слов, отражающих конъюнкцию: «и», «но» и другие.

В саму модель подается множество признаков, например:

- Само слово;
- Его лемма;
- Часть речи;
- Предыдущее и следующее слова, их леммы, части речи;
- Появлялось ли слово-отрицатель в предыдущих 4 словах;
- Стоит ли слово в превосходной или сравнительной степени;
- Синоним, антоним слова по WordNet;
- Слово-родитель в синтаксической структуре;

и так далее.

После применения модели получается высказывание, в котором помечены отрицательные, положительные слова, а также слова (фразы), отображающие аспект. Далее авторы собирают суммаризацию из этой классификации тем же методом, что и в работе [3]: идентифицируют ближайшее к аспекту слово (фразу), выражающую какой либо сентимент, и выдает как готовую пару «аспект-сентимент».

Одним из самых перспективных подходов решения задачи аспектного-сентимент анализа, основанных на transfer-learning, является работа [2]. Авторы замечают, что почти нет нейросетевых архитектур, которые решают обе подзадачи аспектного сентимент анализа, и также выдвигают предположение, что решение обеих этих задач в контексте одной модели повысит точность конечного результата. Таким образом была предложена архитектура под названием Local Context Focus — Aspect Term Extraction Polarity Classification (LCF-АТЕРС). Ее схема представлена на рисунке 5. Авторы переработали структуру стандартной BERT, дополнив ее некоторыми слоями. Сначала предложение подается в две модели BERT: одна из них будет получать свойства локального контекста, а другая - глобального (авторы допускают возможность использования для этих целей одну BERT'у, для экономии ресурсов). Назовем выход из локальной BERT как  $O_{BERT^l}$ , а из глобальной -  $O_{BERT^g}$ . Далее локальные признаки, полученные с помощью BERT, подаются в разработанный авторами модуль, называемый local context focus. Этот модуль основан на понятии семантического относительного расстояния (SRD), которое позволит понять, находится ли слово в радиусе выделенного аспекта. Вычисляется SRD как:  $SRD_i = |i - P_a| - \lfloor \frac{m}{2} \rfloor$ , где  $i$  - индекс анализируемого токена,  $P_a$  - центральная позиция аспекта,  $m$  - длина аспекта, и  $SRD_i$  представляет семантическое расстояние между анализируемым токеном и аспектом.

Далее, на основе этого расстояния, можно подать выход из локальной BERT в один из модулей:

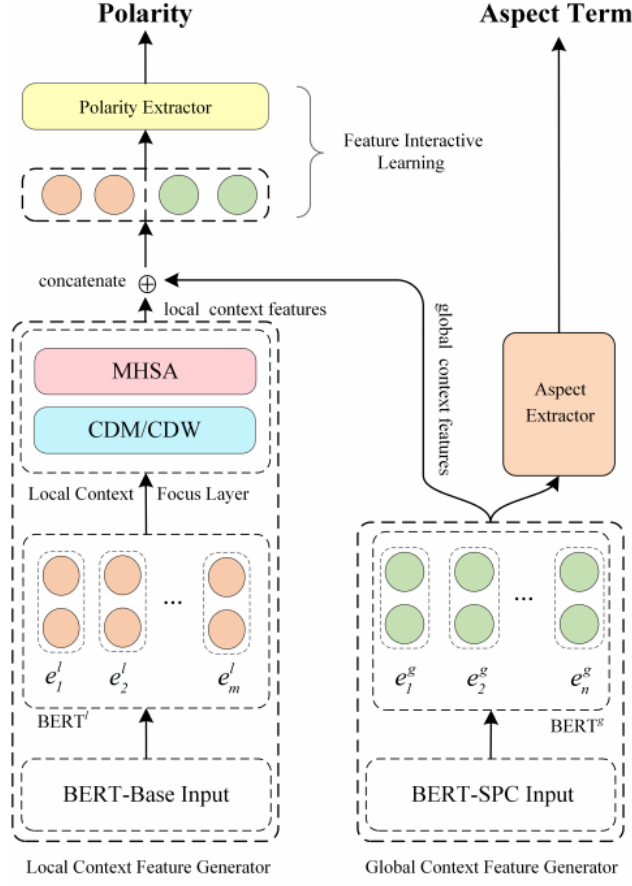


Рис. 5: Архитектура LCF-ATEPC.

- *Динамическое маскирование на основе контекстных признаков (CDM)*: если  $SRD_i$  токена больше определенного порога, то эмбединг этого токена домножится на ноль:

$$V_i = \begin{cases} E & \text{if } SRD_i \leq \alpha \\ 0 & \text{if } SRD_i > \alpha \end{cases}$$

$$V = [V_1, \dots, V_n], O_{CDM}^l = O_{BERT^l} \cdot V$$

- *Динамическое взвешивание на основе контекстных признаков (CDW)*: множитель, на который домножится эмбединг  $i$ -того токена, определяется из формулы:

$$V_i = \begin{cases} E & \text{if } SRD_i \leq \alpha \\ \frac{n - (SRD_i - \alpha)}{n} \cdot E & \text{if } SRD_i > \alpha \end{cases}$$

$$V = [V_1, \dots, V_n], O_{CDW}^l = O_{BERT^l} \cdot V$$

- *Слияние маскирования и взвешивания (Fusion)*: выходы с двух вышеперечисленных модулей конкатенируются и линейно преобразуются:  $O_{fusion}^l = [O_{CDM}^l; O_{CDW}^l], O_{fusion}^l = W \cdot O_{fusion}^l + b$

После этого выход, из какого бы модуля он не получился, обрабатывается механизмом Multi-Head Self-Attention.

После этого выходы из глобальной BERT и Multi-Head Self-Attention конкатенируются и снова пропускаются через линейное преобразование и Self-Attention. Далее, происходит классификация с помощью функции Softmax.

Модель LCF-ATEPC показывает state-of-the-art результаты.

### 5.3 Машинное обучение без учителя

Авторы метода [10] расширили свое исследование в статье [19], посвященное задаче АТЕ, включив в модель решения последней набор классификаторов для каждого аспекта, чтобы искать его сентимент. Проблема в том, что некоторые аспекты могут коррелировать между собой, и вообще соотноситься с общим настроением обзора. Поэтому модель также должна учитывать и связи между аспектами.

Еще одно интересное исследование было проведено [20], была представлена модель, которая выделяет аспекты и сентимент для отзывах о ресторанах. Одно из нововведений тут является то, что в зависимости от сущности (то есть ресторана) набор слов (фраз), в которых модель будет искать аспекты, будет меняться, но распределение слов настроения остается тем же. Далее, используются НММ для моделирования того факта, что слова-аспекты и слова-сентименты могут появляться в определенном порядке. Также, глобально моделируется распределение фоновых слов, которые могут мешать принятию решений.

### 5.4 Сравнение методов

Очевидно, что данный раздел методов является самым универсальным, потому что для использования (не обучения) этой модели не нужно будет иметь аспекты, размеченные вручную или другой моделью. Синтаксические модели все еще очень хороши для отзывов в социальных сетях по уже описанным выше причинам. Не нейросетевые подходы машинного обучения с учителем, как видно, требуют много метаинформации, как, например, доступ к WordNet. Описанные методы машинного обучения без учителя, как видно, работают на задаче topic-modeling, что не очень подходит для коротких текстов.

Нужно выделить модель LCF-АТЕРС как одну из самых перспективных моделей, как будет видно ниже. Она работает на основе BERT, которая уже показала свою мультилингвальность и возможность хорошей обработки коротких текстов.

Далее представлена сравнительная таблица методов, описанных выше.

Авторы метода	Класс метода	Метрики качества
[17]	Syntax-based	precision: 94.3%, recall: 28.6%
[18]	Supervised machine learning	precision: 82.6%, recall: 76.2%
[2]	Supervised machine learning	accuracy: 90.18%
[19]	Unsupervised Machine Learning	precision: 74.5%
[20]	Unsupervised Machine Learning	precision: 74.3% recall: 86.3%

## 6 Результаты работы

Был проведен обзор одних из самых перспективных методов решения задачи аспектного сентимент-анализа. Приведены сравнительные таблицы по решению каждой из подзадач и всей задачи целиком.

Однако был выявлен тот факт, что сравнение методов не максимально «честно»: авторы методов во многих случаях собирали данные для тестирования самостоятельно и поэтому все методы замерялись на разных датасетах. Следовательно, дальнейшие перспективы развития понятны: необходимо реализовать предложенные методы и анализировать их на одном датасете.

## Список литературы

- [1] K. Schouten and F. Frasincar, «Survey on Aspect-Level Sentiment Analysis,» in *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 3, pp. 813-830, 1 March 2016, doi: 10.1109/TKDE.2015.2485209.
- [2] Heng Yang, Bqing Zeng, Jianhao Yang, Youwei Song, Ruyang Xu, «A multi-task learning model for Chinese-oriented aspect polarity classification and aspect term extraction», *Neurocomputing*, Vol. 419, 2021, Pages 344-356, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2020.08.001>.
- [3] Mingqing Hu and Bing Liu. 2004. «Mining opinion features in customer reviews». In *Proceedings of the 19th national conference on Artificial intelligence (AAAI'04)*. AAAI Press, 755–760.
- [4] Rakesh Agrawal and Ramakrishnan Srikant. 1994. «Fast Algorithms for Mining Association Rules in Large Databases». In *Proceedings of the 20th International Conference on Very Large Data Bases (VLDB '94)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 487–499.
- [5] Zhen Hai, Kuiyu Chang, and Jung-jae Kim. 2011. «Implicit feature identification via co-occurrence association rule mining». In *Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I (CICLing'11)*. Springer-Verlag, Berlin, Heidelberg, 393–404.
- [6] C. Scaffidi, K. Bierhoff, E. Chang, M. Felker, H. Ng, and C. Jin, «Red opal: Product-feature scoring from reviews,» in Proc. 8th ACM Conf. Electron. Commerce, 2007, pp. 182–191
- [7] Yanyan Zhao, Bing Qin, Shen Hu, and Ting Liu. 2010. «Generalizing syntactic structures for product attribute candidate extraction». In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (HLT '10)*. Association for Computational Linguistics, USA, 377–380.
- [8] Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. 2009. «Expanding domain sentiment lexicon through double propagation». In *Proceedings of the 21st international joint conference on Artificial intelligence (IJCAI'09)*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1199–1204.
- [9] Niklas Jakob and Iryna Gurevych. 2010. «Extracting opinion targets in a single- and cross-domain setting with conditional random fields». In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP '10)*. Association for Computational Linguistics, USA, 1035–1045.
- [10] Ivan Titov and Ryan McDonald. 2008. «Modeling online reviews with multi-grain topic models». In *Proceedings of the 17th international conference on World Wide Web (WWW '08)*. Association for Computing Machinery, New York, NY, USA, 111–120. DOI:<https://doi.org/10.1145/1367497.1367513>
- [11] Lakkaraju, Hima Bhattacharyya, Chiranjib Bhattacharya, Indrajit. (2011). «Exploiting Coherence for the Simultaneous Discovery of Latent Facets and associated Sentiments.» 498-509. 10.1137/1.9781611972818.43.
- [12] Samaneh Moghaddam and Martin Ester. 2010. «Opinion digger: an unsupervised opinion miner from unstructured product reviews». In *Proceedings of the 19th ACM international conference on Information and knowledge management (CIKM '10)*. Association for Computing Machinery, New York, NY, USA, 1825–1828. DOI:<https://doi.org/10.1145/1871437.1871739>
- [13] Yejin Choi and Claire Cardie. 2008. «Learning with compositional semantics as structural inference for subsentential sentiment analysis». In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP '08)*. Association for Computational Linguistics, USA, 793–801.
- [14] Xin Li, Lidong Bing, Wai Lam, and Bei Shi. 2018. «Transformation networks for target-oriented sentiment classification». In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 946–956.
- [15] Song, Youwei Wang, Jiahai Jiang, Tao Liu, Zhiyue Rao, Yanghui. (2019). «Attentional Encoder Network for Targeted Sentiment Classification.»

- [16] Ana-Maria Popescu and Oren Etzioni. 2005. «Extracting product features and opinions from reviews.» *In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT '05)*. Association for Computational Linguistics, USA, 339–346. DOI:<https://doi.org/10.3115/1220575.1220618>
- [17] Tetsuya Nasukawa and Jeonghee Yi. 2003. «Sentiment analysis: capturing favorability using natural language processing.» *In Proceedings of the 2nd international conference on Knowledge capture (K-CAP '03)*. Association for Computing Machinery, New York, NY, USA, 70–77. DOI:<https://doi.org/10.1145/945645.945658>
- [18] Fangtao Li, Chao Han, Minlie Huang, Xiaoyan Zhu, Ying-Ju Xia, Shu Zhang, and Hao Yu. 2010. «Structure-aware review mining and summarization.» *In Proceedings of the 23rd International Conference on Computational Linguistics (COLING '10)*. Association for Computational Linguistics, USA, 653–661.
- [19] Titov, Ivan, McDonald, Ryan,. (2008). «A Joint Model of Text and Aspect Ratings for Sentiment Summarization.»
- [20] Christina Sauper and Regina Barzilay. 2013. «Automatic aggregation by joint modeling of aspects and values.» *J. Artif. Int. Res.* 46, 1 (January 2013), 89–127.