Эксперименты по выделению тональных слов проводились на данных корпуса с отзывами на рестораны. Корпус включал в себя часть из N текстов, в которой по оценке и аспекту был размечен каждый текст и размеченную нами часть из 11 текстов, в которой по аспекту и оценке был размечен каждый токен. Первоначальный замысел состоял в том, чтобы сначала обучить классификатор (регуляризованную логистическую регресию) для категоризации текстов по аспектным оценкам, затем использовать этот классификатор для выделения тональных слов следующим образом: 1) для каждого слова в предложении выделить все контексты фиксированной длины, в которые оно входит; 2) для каждого контекста оценить классификатором принадлежность к положительному/отрицательному толнальному классу. Если вероятность оказывалась выше заданного порога, слово считалось бы положительно/отрицательно тональным. Подход подразумевал три настраиваемых параметра – длину контекста, пороговое значение, а также, то какую из вероятностей среди контекстов – минимальную, максимальную или среднюю – использовать для сравнения с пороговым значением. Параметры подбирались при помощи сравнения с ручной разметкой с использованием F1-меры. Результаты применения такого подхода оказались безуспешными: при большинстве значений настраиваемых параметров пересечение с ручной разметкой по F1-мере было близко к нулю, а лучший результат составил около 0.2. С применением данного подхода можно ознакомиться в ноутбуке <https://github.com/nicklogin/NLPTeam3000/blob/master/find_tonal_words.ipynb>.

Следующим решением было воспользоваться подходом, используемым в реферируемой статье (Wang, Y., Zhang, Y. and Liu, B., 2017) и выделять тональные слова вне контекста в рамках задачи PU (Positive/Unlabeled) Learning – использовать размеченную и неразмеченную выборку, часть тональных слов взять за множество «слов-шпионов» и добавить к неразмеченной выборке, обучать классификатор слов до тех пор пока он не сможет извлечь из неразмеченной выборки все слова-шпионы, сохранить новые слова из неразмеченной выборки, которые выделит классификатор. В качестве размеченного множества было решено использовать словарь русской тональной лексики проекта ЛИНИС (<http://linis-crowd.org/>). В качестве входной информации для классификатора предполагается использовать эмбеддинги слов. В качестве модели векторного представления слов нами использовалась модель FastText SkipGram, обученная на лемматизированных текстах Национального русского языка из репозитория RusVectores (<http://vectors.nlpl.eu/repository/11/181.zip> ). Однако так как задача подразумевала выделение именно характерной для аспектов Еда/Сервис лексики было решено сначала найти наиболее близкие к выделенным слова среди размеченных как отрицательные/положительные в датасете ЛИНИС и уже их использовать как размеченные принадлежащие к искомому классу примеру. В качестве меры расстояния использовалось косинусное расстояние между эмбеддингами (для биграмм и триграмм – использовались взвешенные усреднённые эмбеддинги, где в качестве весов выступали значения IDF). Однако при подобранных так, чтобы в полученный сет не попали слова-аномалии, нами пороговых значениях расстояния общего количества ранее выделенных и новых тональных слов (129) оказалось слишком мало по сравнению с объёмом используемого размеченного сета (7546), поэтому решено было остановиться на данном этапе – выделению из размеченного сета по косинусной близости.

Похожим образом было решено выделить тональные слова в development-корпусе. Для начала корпус был предобработан (лемматизирован и токенизирован, синтаксические связи не использовались при помощи парсера Udpipe). Биграммы и триграммы из development-корпуса, используя единицы, частота которых была бы выше 5, при этом слова и N-граммы для сравнения с положительными/отрицательными списками по Еде/Обслуживания извлекая из текстов с оценкой по соответствующему параметру ниже 6/6 и выше. Такой подход помог определить в development-корпусе 361 положительно-тональную в отношении Еды конструкцию, 28 отрицательно-тональных, 211 положительно-тональных и 14 отрицательно-тональных в отношении сервиса. При пересечении с ранее выдленными словами и биграммами мы получили такой рездультат (в клетках таблицы отражено количество найденных единиц):

|  |  |  |
| --- | --- | --- |
| Аспект/Оценка | Положительная | Отрицательная |
| Еда | 365 | 50 |
| Сервис | 218 | 26 |

Вышеописанную процедуру выделения тональных слов можно найти в ноутбуке: <https://github.com/nicklogin/NLPTeam3000/blob/master/find_tonal_words.ipynb>

Несколько странных случаев выделения оченочной лексики были исправлены в автоматически сгенеированном тональном словаре вручную.

Скрипт, выделяющий тональные слова в CONLLU-разметке использует списки слов и N-грамм, полученные при помощи вышеописанной процедуры.