Санкт-Петербургский государственный университет
Филологический факультет
Кафедра математической лингвистики

Васильева Анна Станиславовна

Проблема разработки системы оценки тональности сообщений на украинском языке

Выпускная квалификационная работа по направлению
45.03.02 «Лингвистика»,
образовательная программа «Прикладная, экспериментальная и математическая лингвистика»

Научный руководитель:
доц., к.ф.н. Митренина О.В.

Санкт-Петербург
2017

Оглавление

Введение

Анализ тональности текста (сентимент-анализ, англ. Sentiment analysis) — класс методов анализа текста в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и эмоциональной оценки авторов относительно объектов в тексте [58].

Мнение окружающих на протяжении многих веков влияло на различные сферы деятельности человека. Однако с распространением интернета это влияние значительно укрепилось. Раньше людям предоставлялась возможность узнать мнение лишь у ограниченного числа собеседников. Теперь же с появлением интернет-магазинов, блогов, социальных сетей, специализированных ресурсов («Яндекс.Маркет», «Epinions.com», «Кинопоиск») пользователи могут обращаться за мнением к большой аудитории.

Крупные компании и организации также активно используют подобные ресурсы для исследования конкурентной среды, наблюдения за состоянием рынка с целью его оценки.

Социальные сети предоставляют исследователям широкое поле для проведения детального анализа мнений пользователей. К примеру, американский проект Pulse of the Nation [61] был создан для того, чтобы в течение дня отслеживать настроение граждан, пользующихся соцсетью Twitter.

«Твиттер» (Twitter) — одна из самых популярных социальных сетей для публичного обмена сообщениями. По состоянию на февраль 2016 года сервис насчитывает около 305 млн активных пользователей. Сообщения настроены на 140 символов для совместимости с SMS-сообщениями.

Целью данной работы является выявление и анализ проблем, связанных с разработкой системы оценки тональности текстов на украинском языке на примере системы SentiStrength. Программа SentiStrength, созданная как часть проекта CyberEmotions, автоматически производит анализ тональности коротких текстов. Она основана на использовании словаря эмоциональной лексики и корректирующих правил.

Для достижения поставленной цели перед нами были поставлены следующие теоретические и практические **задачи**:

1) изучить применение, задачи и проблемы анализа тональности, а также основные понятия, связанные с ним;
2) рассмотреть основные подходы для решения задач сентимент-анализа;
3) описать основные принципы работы инструмента SentiStrength для анализа тональности;
4) провести настройку программы SentiStrength на украинский язык;
5) оценить эффективность работы программы для украинских текстов.

В работе мы использовали **методы** машинного обучения и анализа тональности с использованием словарей эмоциональной и оценочной лексики также. Словарь эмоциональных слов украинского языка создавался с помощью экспертов и автоматически. **Материалом** исследования стала случайная выборка коротких текстов на украинском языке из социальной сети твиттер объёмом 1200 сообщений.

Данная работа имеет большую **практическую значимость**, так как результаты настройки программы SentiStrength могут быть использованы разработчиками инструмента, что позволит исследователям аудитории носителей украинского языка, а также различным компаниям и организациям использовать данный продукт для анализа текстов на украинском языке.

Новизна работы определяется тем, что в мире пока не существует доступных систем автоматической оценки тональности текстов на украинском языке.

Работа состоит из введения, четырёх глав, заключения, списка литературы и двух приложений. В первой главе даётся подробный обзор предметной области. Во второй главе рассматриваются различные методы определения тональности текстов. В третьей главе подробно описывается работа программы SentiStrength, а также приводятся предпосылки её создания. В четвертой главе работы описывается

процесс настройки программы на украинский язык и оценена эффективность системы.

Глава 1. Обзор предметной области

1.1 Применение анализа тональности

Мнение — это центральное понятие практически любой деятельности человека. Мнения являются одним из ключевых источников влияния на наше поведение. Всякий раз, когда человек принимает решение, для него важно знать, что об этом думают другие люди. Компании и организации всегда интересуются мнением покупателей и клиентов об их продуктах и услугах. Отдельным покупателям также интересно знать мнение пользователей товара перед его покупкой. Многие люди стремятся узнать, что думает общественность по поводу того или иного политического кандидата перед голосованием на выборах. В прошлом, когда человек нуждался в чьём-то мнении, он спрашивал его у членов своей семьи и у друзей. Компании проводили опросы и анкетирования.

Вместе с огромным развитием социальных сетей (сайтов с отзывами, форумов, блогов, микроблогов, Twitter и др.) в Интернете, люди стали гораздо чаще обращаться к подобным ресурсам, чтобы принимать решения с учётом мнения общественности. Сегодня если человек хочет приобрести какой-либо товар, у него есть возможность узнать мнение не только членов семьи, но ещё и опытных пользователей. Крупным организациям больше не обязательно проводить опросы и анкетирования, так как подобная информация уже в избытке находится в свободном доступе. Однако в связи с огромным количеством сайтов, содержащих отзывы, для обычного пользователя будет весьма проблематичным обобщить огромное множество мнений. Именно для этого и существуют автоматические системы анализа тональности текста.

Анализ тональности используется в таких областях как сфера услуг, здравоохранение, финансовое обслуживание, политические выборы и т.п. Всё это даёт сильную мотивацию для проведения исследований, связанных с сентимент-анализом. Например, в работе [31], модель анализа используется для предсказаний торговых успехов. Исследование [35] было связано с опросами общественного

мнения. В работе [51] мнения из Twitter использовались для предсказания результатов выбора. В исследованиях [7; 24; 42] данные из Twitter о фильмах применялись для предсказания доходов билетных касс.

1.2 Основные понятия

Анализ тональности текста — это область компьютерной лингвистики, посвященная автоматическому выявлению оценок, эмоций человека относительно упоминаемых в тексте сущностей, таких как, например, продукт, услуга, событие, организация, личность и т.п. или выявлению общей оценки, когда просто анализируется эмоциональность высказывания.

В данной работе понятия «анализ тональности текста», «анализ эмоциональной окраски текста» и «сентимент-анализ» используются взаимозаменяемо.

С точки зрения анализа тональности, информация в тексте делится на два класса [2]:

1. Мнения
2. Факты

Как правило, целью анализа тональности текста является выявление мнений и их свойств.

Мнения, в свою очередь, также можно разделить на два типа:

1. Простое мнение (*regular opinion*)
2. Сравнительное мнение (*comparative opinion*).

В простом мнении содержится отношение автора к одному объекту. Мнение может быть выражено как явно (прямое мнение — *direct opinion*), так и неявно (непрямое мнение — *indirect opinion*). Примером прямого мнения является предложение «Качество фотографии отличное», примером мнения, выраженного неявно — «После того, как я выпил это лекарство, мне стало хуже».

Объектом анализа тональности текста является любая сущность (продукт, услуга, проблема, личность и т.п.), относительно которой выражается мнение в тексте. Очень часто, объекты могут состоять из отдельных частей (компонентов) и свойств (*features*). Компоненты и свойства составляют множество аспектов (*aspects*).

Допустим, что объектом мнения является фотоаппарат. Он обладает такими составными частями, как объектив, батарея, дисплей и т.п., а также следующим набором атрибутов: размер, вес, внешний вид и т.д. Мнение может быть выражено как относительно самого объекта, так и относительно его аспектов. Например, в предложении «Хороший фотоаппарат, добротный аккумулятор (быстро заряжается и околомедленно садится)» первая его часть выражает положительное мнение об объекте «фотоаппарат», а вторая — о его аспекте «аккумулятор».

Для удобства объект и его компоненты можно представить иерархически (рисунок 1).

Рисунок 1 – Иерархическое представление объекта и его компонентов

Автор мнения (*opinion holder*) — это человек, выражающий своё мнение. Вместо понятия «автор мнения» также используется понятие «источник мнения» (*opinion source*).

Тональностью или сентиментом (*opinion orientation, sentiment*) называют эмоциональную окраску, выраженную в тексте. Можно выделить тональность трёх типов: положительную, отрицательную и нейтральную. Однако в научной литературе нейтральную тональность определяют по-разному, поэтому очень часто принимают во внимание только первые два типа тональности.

Некоторые исследователи определяют нейтральную тональность как некоторое промежуточное положение между положительной и отрицательной, а некоторые определяют её как отсутствие как какой-либо эмоциональной окраски.

Таким образом, в анализе эмоциональной окраски текста простое мнение определяется через кортеж, состоящий из пяти элементов (entity, aspect, sentiment value, holder, time), где entity — это сущность, об аспекте (aspect) которого автор (holder) выразил мнение (sentiment value) в определённый момент времени (time).

Например, на рисунке 2 изображено сообщение, в котором «@gbolotoff» — автор, «6 дек. 2014г» — дата публикации, «Homeland» - объект мнения, «сюжет» - аспект объекта, а «восхитителен» - слово, которое говорит о том, что мнение автора об аспекте является положительным.

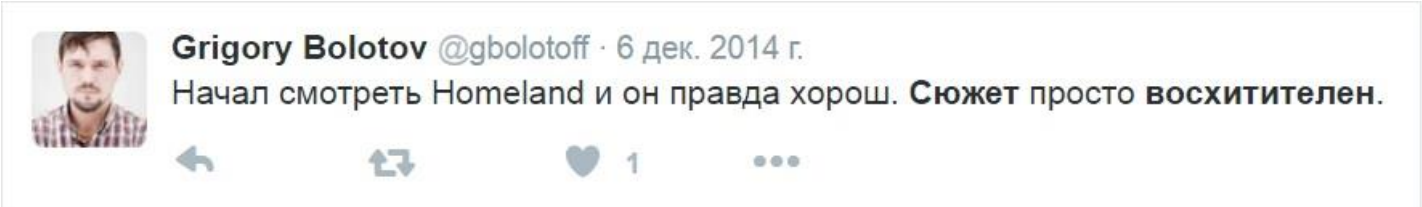

Рисунок 2 – Пример выражения прямого мнения в сообщении из социальной сети Twitter

Второй тип мнений — сравнительный — имеет три подтипа:

1. Сравнение аспектов объектов в пользу одного (*Non-equal gradable comparison*). Оно выражает отношения «что-то лучше/хуже чего-то», которые ранжируют сущности в соответствии с предпочтениями автора (чаще всего используется сравнительная степень сравнения прилагательных). Примером является предложение «Кока-кола вкуснее, чем пепси». Этот подтип включает в себя также предпочтения: «Я предпочитаю пить колу, нежели пепси».
2. Приравнивание аспектов разных объектов (*Equative comparison*). Оно выражает отношение тождественности, которое гласит, что две или более сущностей являются равными на основании сравнения некоторых общих аспектов. Например, «На вкус пепси и кола одинаковы».

3. Превосходство одного объекта над другими (*Superlative comparison*). Этот подтип выражает то же отношение «что-то лучше/хуже чего-то», что и первый, но в данном случае одна сущность располагается над остальными, т.е. в основном используется превосходная степень прилагательных: «Кола — самый вкусный из безалкогольных напитков».

Сравнительные мнения (рисунок 3) определяются следующим образом. Это кортеж (entity$_1$, entity$_2$, aspect, preferred_objects, holder, time), где entity$_1$ и entity$_2$ — это объекты (их может быть больше), сравниваемые по общему аспекту (aspect), preferred_objects — множество объектов, которые автор (holder) предпочёл в момент времени (time).

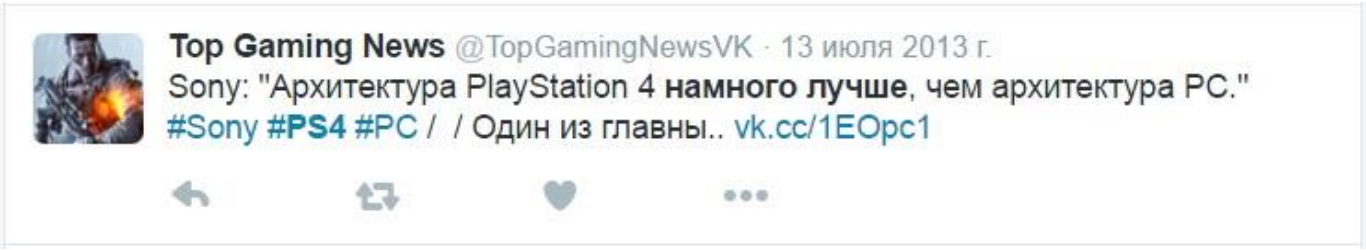

Рисунок 3 – Пример твита, содержащего сравнительное мнение

В примере на рисунке 3 кортеж мнения выглядит так: ({PlayStation_4}, {PC}, {архитектура}, {PlayStation_4}, {Sony}, {13_июля_2013г.}).

Кортеж сравнительного мнения, в отличие от прямого, не содержит оценку эмоций автора.

В анализе тональности текста можно также встретить понятие, тесно связанное с понятием мнения — субъективность.

Согласно работе [30] объективное предложение предоставляет собой некую фактическую информацию о мире, тогда как субъективное предложение выражает чувства, взгляды, убеждение человека. Примером объективного предложения является предложение «iPhone является продуктом компании Apple», а примером субъективного — «Мне нравится iPhone».

Таким образом, пояснив необходимые нам термины, мы можем перейти к описанию задач автоматического анализа тональности текста.

1.3 Задачи анализа тональности

С точки зрения компьютерной лингвистики текст на естественном языке представляет собой неструктурированную информацию. Выше мы дали определение мнения и, таким образом, определили, как информация в тексте должна быть структурирована.

Учитывая, что мнение — это кортеж, можно сформулировать 6 основных задач анализа тональности текста:

Задача 1 (извлечение объектов и их классификация): Необходимо извлечь все слова, выражающие сущности в тексте d, синонимичные слова сгруппировать в классы (или кластеры). Каждый кластер соответствует одной уникальной сущности e_i.

Задача 2 (извлечение аспектов и их классификация): Необходимо извлечь все слова, выражающие аспекты, и синонимичные единицы аналогично распределить по классам. Каждый такой «аспектный» кластер сущности e_i соответствует одному уникальному аспекту a_{ij}.

Задача 3 (извлечение автора и классификация): Из текста или из структурных данных необходимо извлечь слова, выражающие автора мнения, и также классифицировать их. Это задание аналогично первым двум.

Задача 4 (извлечение времени и его стандартизация): Необходимо также извлечь время, когда было написано сообщение, и привести разные его форматы к одному. Эта задача также аналогична тем, что описаны выше.

Задача 5 (определение полярности): Необходимо определить, является ли мнение об аспекте a_{ij} положительным, отрицательным или нейтральным, или присвоить соответствующее численное значение (которое оговаривается заранее).

Задача 6 (вывод кортежа): Последняя задача состоит в том, чтобы породить итоговый кортеж (e_i, a_{ij}, s_{ijkl}, h_k, t_l) мнения, выраженного в документе d, на основании результатов предыдущих задач. На первый взгляд кажется, что эта

задача является очень простой, однако на примере [29], представленном ниже, можно показать обратное:

Posted by: bigJohn Date: Sept. 15, 2011

(1) I bought a Samsung camera and my friends brought a Canon camera yesterday. (2) In the past week, we both used the cameras a lot. (3) The photos from my Samy are not that great, and the battery life is short too. (4) My friend was very happy with his camera and loves its picture quality. (5) I want a camera that can take good photos. (6) I am going to return it tomorrow.

Сначала программа должна выделить из текста следующие сущности: «*Samsung», «Canon», «Samy»* и объединить соответственно в одну группу «*Samsung»* и «*Samy»* и взять отдельно «*Canon»* (задача 1). Затем программа должна извлечь следующие аспекты: *«picture», «photo»* и *«battery life»* и объединить вместе *«picture»* и *«photo»,* так как для камеры это одно и то же (задача 2). Затем программа должна определить, что в предложении (3) источником мнения является автор — *«bigJohn»,* а в предложении (4) — друг Джона — *«bigJohn's friend»* (задача 3). Затем необходимо определить время публикации отзыва — *«Sept. 15, 2011»* (задача 4). В соответствии задачей 5, программа должна определить, что в предложении (3) даётся отрицательное мнение относительно качества фотографий и срока службы аккумулятора камеры Samsung, а в предложении (4) даётся положительный отзыв о камере Canon в целом, а также о качестве её фотографий. Может показаться, что в предложении (5) даётся положительный отзыв, однако это не так. Чтобы сгенерировать кортеж мнения для предложения (4), программа также должна определить, к чему относятся следующие единицы: «his camera» и «its». В конце концов должны получиться следующие кортежи:

(Samsung, picture_quality, negative, bigJohn, Sept-15-2011)

(Samsung, battery_life, negative, bigJohn, Sept-15-2011)

(Canon, GENERAL, positive, bigJohn's_friend, Sept-15-2011)

(Canon, picture_quality, positive, bigJohn's_friend, Sept-15-2011)

Задача 5 (определение полярности) обычно рассматривается как задача текстовой классификации. В связи с этим есть два типа её решения [66]:

1) Использовать «плоскую» (*flat*), т.е. одноуровневую классификацию. В этом случае у нас есть 3 класса (положительный, отрицательный, нейтральный), и из них необходимо выбрать один. Пример плоской классификации изображён на рисунке 4.

Рисунок 4 – Плоская классификация

2) Второй способ — иерархическая классификация. Здесь мы сначала решаем, документ является объективным или субъективным, и если документ субъективный, смотрим, является он положительным или отрицательным (уровней может быть больше, см. рисунок 5)

Рис. 5 Иерархическая классификация

Стоит отметить, что определение полярности может проводиться на нескольких уровнях:

1) на уровне документа (*document-level sentiment classification*);
2) на уровне предложения (*sentence-level sentiment classification*);
3) на уровне аспекта (*feature-level sentiment classification*).

Так как большинство исследователей фокусируют своё внимание на простых (не сравнительных) мнениях, для решения перечисленных выше задач делается следующее допущение: в документе d, который содержит мнение, автор h выражает мнение относительно одной сущности e. Хотя теоретически автор может выражать мнение о нескольких сущностях, в случае отзывов (а особенно для сообщений Twitter) в целом это допущение справедливо.

1.4 Проблемы автоматического определения тональности

К проблемам, возникающим при автоматическом анализе тональности текстов, можно отнести следующие случаи:

1. Положительные и отрицательные слова могут иметь противоположную тональность в текстах на разные темы, т.е. анализ сильно зависит от предметной области. Например, фраза «почитайте книгу» может подразумевать положительную оценку в отзыве о книге, но отрицательную в рецензии на фильм; слово «непредсказуемый» может выражать положительные эмоции относительно фильма, но отрицательные относительно сферы обслуживания.

   Как показывает практика, интересы пользователей не ограничиваются одной лишь предметной областью, поэтому возможно решение данной проблемы в два этапа [4]: сначала провести тематическую классификацию документа, а затем уже классификацию тональности.
2. Предложения, которые содержат эмоциональную лексику, могут не выражать никаких эмоций, т.е. могут являться объективными. Примерами могут служить вопросительные и условные предложения: «Можете посоветовать какую-нибудь интересный фильм?» или «Если я найду в

магазине хорошую камеру, я её куплю». Однако не все условные и вопросительные предложения попадают в эту группу, т.е. они могут быть и субъективными: «Кто-нибудь знает, как починить этот ужасный принтер?» или «Если вы ищете хорошую машину, купите Toyota Camry».

3. Предложения с сарказмом (содержащие или не содержащие эмотивную лексику [3]) плохо поддаются автоматическому анализу. В высказываниях, содержащих сарказм, общая тональность может иметь противоположное значение отдельных *эмоциональных слов*: «Что за потрясающая машина! Перестала работать через два дня после покупки». Саркастические высказывания не так часто встречаются в отзывах о продуктах и услугах, но довольно употребительны в политических дискуссиях, что делает сложной обработку мнений о политике.

   В одной из работ по анализу сообщений, содержащих сарказм, исследователям удалось добиться точности в 78% на выборке отзывов на товары [33]. Авторы использовали метод частичного обучения с учителем (*semi-supervised learning*).
4. Значение тональности также зависит от пользователя, для которого важно мнение. Например, для компании Apple в предложении «Новый iPhone 6 продаётся на ура ☺» будет положительная тональность, тогда как для компании Samsung — отрицательная.
5. Многие предложения, не содержащие эмотивную лексику, также могут подразумевать мнения. На самом деле большинство таких предложений являются объективными, они выражают некую фактическую информацию. Например, предложение «Эта посудомоечная машина использует много воды» в неявном виде выражает отрицательную оценку о машине, так как она затрачивает много ресурса (воды). Предложение «Я поспал на матрасе два дня, после чего в центре образовалась впадина» также выражает отрицательное мнение относительно матраса. Однако стоит ещё раз отметить, что эти предложения являются объективными.

1.5 Выводы к главе 1

Анализ тональности текста — это область компьютерной лингвистики, посвященная автоматическому выявлению оценок, эмоций человека относительно сущностей в тексте или выявлению общей оценки эмоциональности высказывания. Тональность — это эмоциональная окраска, выраженная в тексте. Данный анализ применяется как в коммерческих целях, так и для решения научно-исследовательских задач.

Различают два типа мнений — простые и сравнительные. Большинство работ, посвящённых данной области исследования, занимаются выявлением простых мнений, так как сравнительные крайне тяжело поддаются анализу.

Чаще всего выделяют 6 задач анализа эмоциональной окраски текста — извлечение объектов, аспектов, автора и их классификация; извлечение времени и его стандартизация; определение тональности; вывод мнения.

К основным проблемам, которые возникают при анализе тональности, можно отнести зависимость тональности от предметной области, использование эмотивной лексики в нейтральных предложениях, сарказм, зависимость тональности от пользователя, который читает сообщение, а также выражение тональности без использования эмоционально окрашенных слов. Данные проблемы с разным успехом могут быть устранены в процессе сентимент-анализа.

Глава 2. Методы автоматического определения тональности

2.1 Основные подходы

Основные методы к определению тональности текста можно разделить на следующие группы:

1. **Подход, основанный на правилах** (*rule-based approach).* Такой подход состоит из определённого набора правил, на основании которых система делает вывод о тональности текста. Примером такого правила для предложения «Я купил потрясающий телефон» является следующее утверждение:

Если прилагательное («потрясающий») входит во множество положительных прилагательных («потрясающий», «восхитительный», «хороший», ...) и не входит во множество отрицательных прилагательных («ужасный», «отвратительный», «плохой», …), то классифицировать тональность как «положительная».

Достоинства:

- данный подход может давать хорошие результаты при большом наборе правил

Недостатки:

- составление большого набора правил — **о**чень трудоёмкий процесс
- очень часто правила привязываются к определённой тематической области
- этот подход не очень подходит для анализа микроблогов из-за «зашумлённости» данных, обусловленной наличием ошибок

Применение данного подхода для документов на русском языке более подробно описано в работах [4; 26].

2. **Подход, основанный на использовании словарей эмоциональной и оценочной лексики** (*sentiment lexicon*). В данном случае каждому отдельному слову в словаре присваивается значение тональности (шкалы оговариваются заранее). Для получения итогового значения тональности часто используют простой способ: берут среднее арифметическое или вычисляют сумму значений тональности всех слов из документа. Более сложный способ — обучение классификатора (например, нейронной сети).

   Примерами подобных словарей для английского языка являются SentiWordNet, ANEW и др.

   Достоинства:

   - данный метод прост в применении

   Недостатки:

   - метод не универсален, для новой предметной области требуется составление нового словаря

   На данном подходе основан принцип работы программы SentiStrength. Принципам ее работы посвящена глава 3.

3. **Методы, основанные на машинном обучении с учителем** (s*upervised learning*). Это наиболее распространённый метод. Его суть состоит в том, чтобы обучить алгоритм классификации на основе коллекции документов (выборки), классы которых известны заранее.

   Достоинства:

   - высокая точность при определении тональности
   - проблема зависимости от конкретной предметной области решается путём обучения классификатора на основе выборки из данной области, так как классификатор сам выделяет признаки, которые влияют на тональность
   - проводится множество исследований с целью улучшения точности

   Недостатки:

- необходима размеченная коллекция текстов (разметка является весьма трудоёмким процессом)

4. **Методы, основанные на машинном обучении без учителя** (*unsupervised learning*). Отличие от предыдущего метода заключается в том, что скрытые закономерности и взаимосвязи между объектами выявляются из неразмеченной выборки данных (либо берутся размеченные данные, но такая информация при обучении алгоритма не используется).

   Достоинства:

   - не требуется размеченная коллекция документов

   Недостатки:

   - низкая точность, по сравнению с обучением с учителем

   Рассмотрим более подробно методы обучения с учителем.

2.2 Методы, основанные на обучении с учителем

Алгоритм реализации данного подхода можно кратко описать следующим образом:

1) Сначала необходимо собрать коллекцию документов, на основе которой будет обучаться классификатор;
2) Каждый документ необходимо представить в виде вектора признаков (аспектов);
3) Далее каждому документу нужно присвоить правильный тип тональности;
4) Необходимо выбрать алгоритм классификации и метод для обучения классификатора;
5) Применение полученной модели.

Прежде чем переходить к вышеописанному алгоритму, необходимо решить, какое количество классов и какой тип классификации будут использованы.

При использовании плоской классификации весьма сложно достичь высоких результатов. Исследования [8; 23] показывают, что гораздо лучшие результаты даёт иерархическая классификация.

Все документы из обучающей выборки должны представлять собой n-мерные векторы аспектов. От того, какой набор характеристик будет использован, напрямую зависит качество результатов. Наиболее распространёнными способами представления документов — это либо в виде так называемого «мешка слов» (bag-of-words), либо в виде n-грамм.

- **«мешок слов».** Допустим, есть 2 документа:

  (1) John likes to watch movies. Mary likes movies too.

  (2) John also likes to watch football games.

  На их основе выделяется список встречающихся слов (часто слова приводят в их начальную форму при помощи стемминга):

  ["John", "likes", "to", "watch", "movies", "also", "football", "games", "Mary", "too"]

  В соответствии с этим списком документы можно представить следующим образом:

  (1) (1, 2, 1, 1, 2, 0, 0, 0, 1, 1)

  (2) (1, 1, 1, 1, 0, 1, 1, 1, 0, 0)

  Размерность векторов соответствует размерности списка. Числа показывают то, сколько раз слово встречается в данном документе.

- **n-граммы**. Допустим, есть документ «Я купил потрясающий ноутбук». Набором униграмм в данном случае будет являться список последовательностей («Я», «купил», «потрясающий», «ноутбук»), набором биграмм — («Я купил», «купил потрясающий», «потрясающий ноутбук»). Чаще всего в анализе используются униграммы, биграммы, либо их комбинация: («Я», «купил», «потрясающий», «ноутбук», «Я купил», «купил потрясающий»,

«потрясающий ноутбук»). В задачах текстовой классификации в большинстве случаев используются n-граммы с 1<n<3.

Далее, чтобы составить вектор, необходимо присвоить каждому его признаку вес. В информационном поиске популярным методом оценки веса является мера TF-IDF, однако она не очень эффективна в анализе тональности. В работе [33] используется его модификация — дельта TF-IDF. Суть заключается в том, чтобы присвоить больший вес для слов, которые имеют положительную или отрицательную тональность. Формула расчёта следующая:

$$V_{t,d} = C_{t,d} \cdot log(\frac{|N| \cdot P_t}{|P| \cdot N_t}) \quad (1)$$

где $V_{t,d}$ — вес слова в документе

$C_{t,d}$ — количество раз слово t встречается в документе d;

|P| — количество документов с положительной тональностью;

|N| — количество документов с отрицательной тональностью;

P_t — количество положительных документов, где встречается слово t;

N_t — количество отрицательный документов, где встречается слово t.

Следующим шагом является выбор алгоритма классификации. Рассмотрим 2 самых распространённых метода.

Наивный байесовский классификатор [65] (*naive Bayes classifier*) — один из самых простых методов классификации. Пусть у нас есть строка O, множество классов C, к одному из которых необходимо отнести строку. Мы выбираем класс так, чтобы вероятность принадлежности объекта к нему была максимальна:

$$c = arg \max_C P(C|O) \quad (2)$$

Вычислить P(C|O) сложно, но можно использовать теорему Байеса:

$$P(C|O) = \frac{P(O|C)P(C)}{P(O)} \quad (3)$$

где P(C) — априорная вероятность гипотезы C;

P(C|O) — вероятность гипотезы C при наступлении события O (апостериорная вероятность);

P(O|C) — вероятность наступления события O при истинности гипотезы C;

P(O) — полная вероятность наступления события O.

Так как нам необходим максимум от функции, мы можем отбросить знаменатель (он в данном случае константа). Так как классификатор работает не со всей строкой, а с вектором признаков, можно представить формулу следующим образом:

$$P(C|o_1 o_2 \dots o_n) = P(o_1 o_2 \dots o_n|C)P(C) \quad (4)$$

Здесь мы делаем «наивное» предположение о том, что переменные O зависят только от класса C и не зависят друг от друга. Правая часть уравнения принимает вид:

$$P(o_1 o_2 \dots o_n|C)P(C) = P(C)P(o_1|C)P(o_2|C) \dots P(o_n|C) = P(C)\prod_i (o_i|C) \quad (5)$$

Конечный вид формулы следующий:

$$c = arg\max_{c \epsilon C} P(\mathrm{c}|o_1 o_2 \dots o_n) = arg\max_{c \epsilon C} P(\mathrm{c})\prod_i (o_i|C) \quad (6)$$

Получается, что всё, что нужно сделать — это вычислить вероятности P(C) и P(O|C). Вычисление этих параметров и называется тренировкой классификатора.

Метод опорных векторов [64] (*SVM, support vector machine*). Целью такой классификации является поиск гиперплоскости в пространстве аспектов, разделяющей все объекты на 2 класса. Идею метода удобно представить на следующем примере: допустим, даны точки на плоскости, разделённые на 2 класса (рисунок 6). Проведём линию между классами. Затем все новые точки (не из обучающей выборки) будут автоматически классифицироваться в соответствии с условиями: точка выше прямой попадает в класс A, а точка ниже прямой — в B.

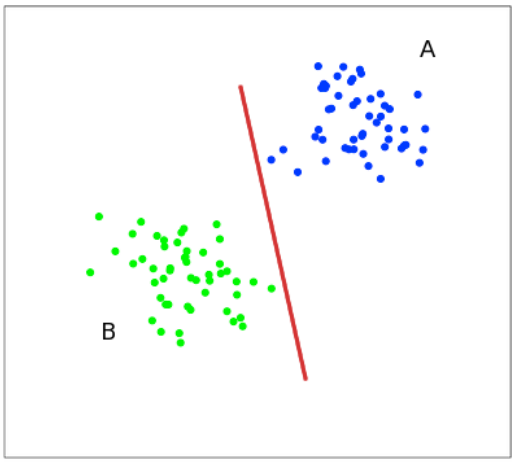

Рисунок 6 – Разбиение точек на классы

Такую прямую можно назвать разделяющей прямой. Но в пространствах высоких размерностей прямая уже не будет разделять классы, поэтому вместо неё рассматривают гиперплоскость — пространство, размерность которого на единицу меньше, чем размерность исходного пространства.

В данном примере существует несколько прямых, разделяющих точки на классы (рисунок 7), однако необходимо выбрать прямую, расстояние от которой до каждого класса максимально. На рисунке 6 это красная прямая.

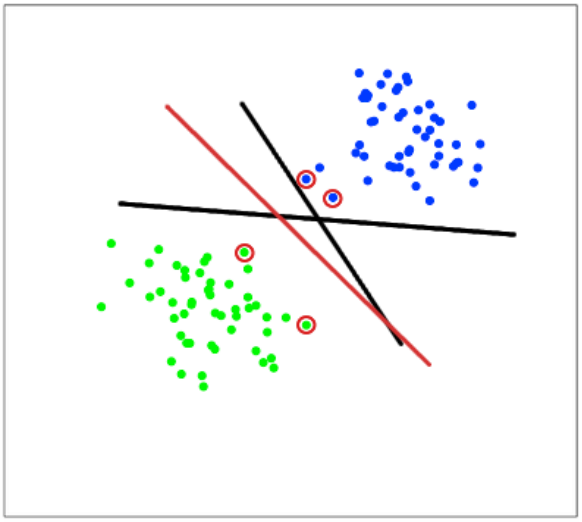

Рисунок 7 – Пример выбора необходимой гиперплоскости

Вектора, лежащие ближе всех к разделяющей плоскости, называются опорными векторами (на рисунке они обведены в красные кружки).

Однако на практике чаще всего встречаются случаи, которые являются линейно не разделимыми (рисунок 8).

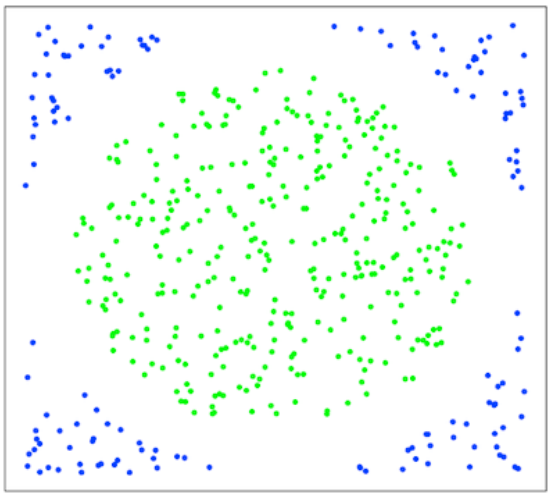

Рисунок 8 – Пример линейно не разделимой выборки

В таких случаях все элементы обучающей выборки вкладываются в пространство более высокой размерности, чтобы в нём выборка была линейно разделима (рисунок 9).

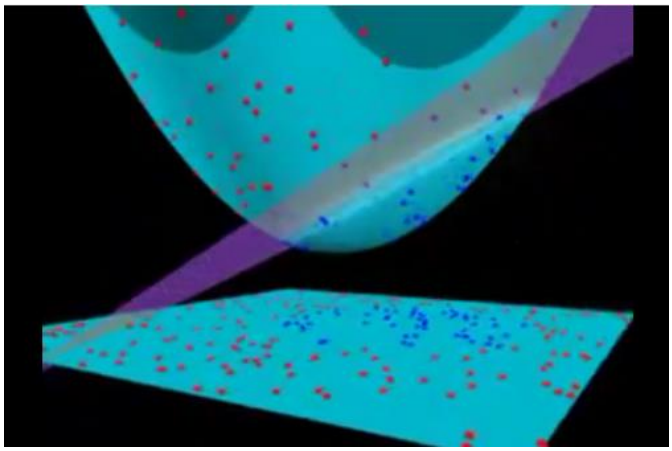

Рисунок 9 – Пример решения проблемы с линейно не разделимой выборкой

После выбора алгоритма классификации и обучения классификатора проводится оценка результатов при помощи полноты и точности, либо при помощи скользящего контроля или перекрёстной проверки (*cross-validation*).

Формула для нахождения полноты:

$$R = \frac{\text{correctly extracted opinions}}{\text{total number of opinions}} \tag{7}$$

где correctly extracted opinions — правильно определённые мнения;
total number of opinions — общее количество мнений (как найденных системой, так и не найденных).

Формула для нахождения точности:

$$P = \frac{\text{correctly extracted opinions}}{\text{total number of opinions found by system}} \quad (8)$$

где correctly extracted opinions —правильно определённые мнения;
total number of opinions found by system — общее количество мнений, найденных системой.

В случае с перекрёстной проверкой, данные разбиваются на k частей, затем на k-1 частях данных производится тренировка модели, а оставшуюся часть используют для тестирования. И так k раз (рисунок 10).

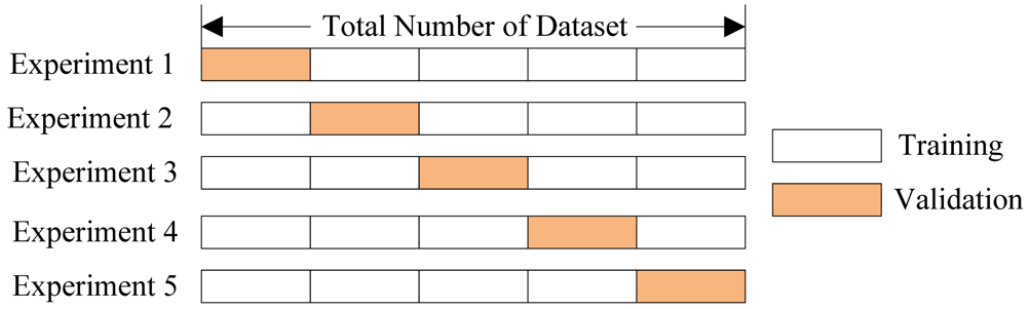

Рисунок 10 – Наглядное представление скользящего контроля

2.3 Выводы к главе 2

Традиционно выделяется 4 основных подхода анализа тональности текстов: метод с использованием правил; метод с использованием словаря эмоциональной лексики; метод, основанный на обучении с учителем, и метод, основанный на обучении без учителя. Каждый из подходов имеет свои достоинства и недостатки.

Подробно был рассмотрен метод, основанный на обучении с учителем, так как он является одним из самых популярных подходов анализа тональности. Данный подход состоит из пяти основных этапов. На первом этапе производится подготовка обучающей выборки. Далее каждый документ представляется в виде вектора признаков (чаще всего это либо «мешок слов», либо n-граммы) с

дальнейшим присваиванием весов каждому элементу вектора. Затем производится выбор алгоритма классификации. В данной главе мы рассмотрели принципы работы двух наиболее эффективных алгоритмов – наивный байесовский классификатор и метод опорных векторов.

Для оценки работы модели можно либо посчитать полноту и точность, либо произвести перекрёстную проверку.

Глава 3. SentiStrength как инструмент для анализа тональности

3.1. Предпосылки создания SentiStrength

Алгоритмы для определения мнения и силы мнения необходимы для того, чтобы помочь понять, какова роль эмоций в неформальном общении, а также чтобы определить неуместные и аномальные эмоциональные высказывания, которые потенциально связаны с угрожающим поведением человека по отношению к себе и окружающим. Тем не менее, существующие алгоритмы определения мнения по большей части направлены на коммерческое использование, например, созданы для определения мнений о продуктах, а не для исследования человеческого поведения. Новый алгоритм SentiStrength частично заполняет это упущение. Сейчас всё больше возрастает интерес к эмоциональной составляющей текстов из социальных сетей, и в частности, из сети Twitter. Большинство алгоритмов анализа тональности не совсем подходят для данной задачи, так как они используют неявные признаки тональности, которые могут скорее отражать особенности жанра или темы, а не эмоциональный аспект. Как результат, такие неявные признаки могут давать ошибочные результаты для исследования социальных сетей, обнаруживая ложные модели. Программа SentiStrength изначально создана для извлечения силы тональности из неформальных английских текстов, она использует новые методы применения уже существующих грамматик и манеры правописания в киберпространстве.

Фактором, затрудняющим выявление тональности в сети, является существование большого количества средств массовой коммуникации, в которых при текстовом общении очень часто игнорируются правила грамматики и правописания. Примером служит язык текстовых сообщений на мобильных телефонах, в которых используются аббревиатуры, эмотиконы и сокращённые предложения, но очевидно, что такой стиль письма можно найти и в формах коммуникации через другие электронные устройства (например, компьютеры). Наиболее широко признанные методы включают обработку эмотиконов типа «☺»,

что, несомненно, является эффективным средством передачи эмоций [18; 21], а также сокращений слов типа m8 (mate), u (you) для английского языка. Подобные изменения создают проблемы, поскольку типичные лингвистические программы анализа тональности начинают работу с частеречной разметки (например, [14]), которая опирается на нормативное правописание и грамматику, и могут применять правила, которые предполагают как минимум правильное правописание, если не грамматически правильно построенные предложения. Исправление правописания может быть полезным с этой точки зрения, но оно основано на предположении, что отклонения в правописании вероятнее всего являются случайными ошибками [28; 40], и поэтому маловероятно, что подобные системы будут проводить анализ успешно при работе с умышленно ненормативным правописанием. Тем не менее, существует большое количество общих сокращений и новых слов, которые лингвистический алгоритм в принципе мог бы обнаружить. Нелингвистические алгоритмы машинного обучения обычно предсказывают тональность на основании встречаемости уни-, би- или триграмм в документах. Этот подход может плохо работать при анализе неофициальных текстов из-за проблем правописания и изощрённости в эмоциональных выражениях, даже если в наличии имеется огромная обучающая выборка.

Согласно статьям [11; 12] известно, что для социальной сети MySpace, на основе которой разрабатывалась программа SentiStrength, характерно то, что её используют в основном молодые люди, сообщения в ней имеют музыкальную направленность, а также в них можно обнаружить типичные языковые модели[1] неформального общения. Возможно, вследствие этих факторов 95% общедоступных комментариев на английском языке, которыми люди делились со своими друзьями, содержат хотя бы одну сокращённую форму, которая может рассматриваться как отклонение от нормативного языка. Общие характеристики сообщений включают эмотиконы, сокращения, часто встречающиеся в СМС-сообщениях, и использование повторяющихся букв или знаков пунктуации,

[1] В данном случае языковые модели (patterns) – образцы текстовых единиц (слов, словосочетаний, предложений).

которые употреблены для эмфазы (например, «a loooong time», «Hi!!!»). Комментарии обычно короткие — среднее число — 18,7 слов, медиана — 13 слов, 68 символов [46], с преобладанием положительных эмоций [50].

Майк Телвалл, Кеван Бакли и их коллеги из университета Уольверхэмптон представили новый алгоритм, SentiStrength, в котором используется несколько новых методов для одновременного извлечения положительной и отрицательной тональности из коротких неформальных текстов в электронном виде [49]. Программа использует словарь эмоционально окрашенных слов с соответствующими весами, отражающими силу тональности, и использует примеры широко употребляемого ненормативного правописания и других общих текстовых способов выражения тональности. Программа разработана с использованием начальной выборки из 2600 комментариев из сети MySpace, размеченных людьми, и оценена на другой случайной выборке из 1041 комментария из MySpace. Следует обратить внимание, что в некоторых статьях, за исключением статей по психологии эмоций, термин тональность подразумевает деление текстов на положительный, отрицательный и нейтральный классы, тогда как понятие эмоция касается более дифференцированного разграничения (счастливый, грустный, испуганный). Однако в случае с SentiStrength два термина используются как синонимы. Новый вклад их работы состоит в следующем: используется метод машинного обучения для оптимизации весов эмоций; используются методы для извлечения тональности из неформальных текстов, содержащих ненормативные слова с повторяющимися буквами, используется метод, исправляющий правописание. В дополнение к этому, система вводит пятибалльную двойную шкалу для положительной и отрицательной тональности, корпус из 1041 комментариев из MySpace для этой системы и совершенно новую систему оценки тональности текстов, которая объединяет новые и уже существующие методы.

3.2. Методы SentiStrength в системе других подходов к анализу тональности

Как уже было описано в первой главе данной работы, извлечение мнения обычно включает в себя три стадии, хотя некоторые задачи (например, [10]) могут потребовать больше шагов. Во-первых, входной текст делится на куски, например, предложения, и каждый отрезок проверяется на наличие тональности — является ли предложение субъективным или объективным [37]. Во-вторых, субъективные предложения анализируются на наличие тональности. И наконец, может выделяться объект, относительно которого высказывается мнение [22]. Обычно метод извлечения мнения занимается только определением положительной или отрицательной тональности, а не отдельными эмоциями (счастье, удивление), не определяет силу тональности и не определяет одновременно положительные и отрицательные эмоции. Тем не менее, проведённые исследования по извлечению мнения могут помочь при выделении одновременно положительных и отрицательных эмоций благодаря тому, что большинство методов могут быть в теории приспособлены под эту новую задачу. Например, методы анализа словосочетаний могут быть применены для определения и положительных, и отрицательных эмоций даже в рамках одного предложения [16; 56; 57]. Менее часто встречающейся задачей является определение силы положительной или отрицательной тональности в одном тексте.

Общепринятым подходом к анализу тональности является выбор алгоритма машинного обучения и метод извлечения характеристик из текста, а затем обучение классификатора на аннотированном человеком корпусе. Характеристиками обычно являются слова, но также могут являться основы слов или слова с частеречной разметкой, которые также могут быть объединены в биграммы (т.е. два следующих друг за другом слова) или триграммы [38]. Были разработаны также более сложные варианты, использующие, например, автоматическое извлечение характеристик [41].

Альтернативным методом определения тональности является выявление наиболее вероятной среднестатистической тональности для слов в тексте, путём оценки того, как часто они встречаются вместе со словами из заданного изначально списка эмотивной лексики, для которой была однозначно определена тональность (например, «хороший», «ужасный»). Для этого обычно используются поисковые системы для оценки относительной частоты совместной встречаемости [52]. Они опираются на предположение о том, что положительные слова чаще встречаются с положительными словами, чем с отрицательными и наоборот. Для данного метода необходим относительно небольшой входной набор лексических данных. Он может приспосабливаться к различным предметным областям в том смысле, что набор начальных общих ключевых слов может использоваться для порождения нового словаря для каждой области применения. Этот метод с использованием первичного списка слов, по всей видимости, даёт весьма хорошие результаты при обработке различных контекстов и распознаёт слова предметной области, связанные с тональностью, такие как, например, 3G для мобильных телефонов.

Однако подобные методы часто склонны выделять слова, которые ассоциируются с тональностью, но явно её не выражают (такие как «чувствовать», «Ирак» или «поздно»). Такие понятия называют «неявно эмоциональными словами» (с англ. indirect affective words) и отличаются от «явно эмоциональных слов» (direct affective words) [44]. Использование неявно эмоциональных слов является недостатком для некоторых типов исследования выявления тональности из социальных сетей, а также для некоторых коммерческих применений, так как они делают методы зависимыми не только от предметной области, но и от времени (например, 3G, вероятно, уже не является надёжным признаком в положительных отзывах на мобильные телефоны).

Лексические алгоритмы

Лексический подход заключается в использовании уже существующего набора понятий с известной тональностью. Затем применяется алгоритм для предсказания тональности текста на основании встречаемости этих слов.

Словарный метод может быть снабжён другой информацией, например, списком эмотиконов и набором семантических правил, которые, например, справлялись бы с отрицанием [34; 45]. Словарь может быть получен из определённого источника, как, например, словари General Inquirer [43], ANEW[2] [13], SentiWordNet [9] WordNet Affect [44] или словарь LIWC[3] [39]. Существуют различные модификации этого метода, которые включают отрицательные частицы [17], слова, которые усиливают тональность других слов («очень любить», «совершенно ненавидеть») и общие структуры предложений [52]. Более сложный подход заключается в определении характеристик текста, которые могли бы потенциально быть субъективными в некоторых контекстах, и в последующем использовании контекстной информации для того, чтобы решить, являются ли характеристики субъективными в каждом новом контексте [55]. Более того, были разработаны различные методы для улучшения стандартных ресурсов, такие как обнаружение сложных слов. Как и описанные выше методы с изначальным списком слов, лексический подход также может находить слова, связанные с тональностью в определённой предметной области, например, слово «маленький» является в общем положительным словом для обзоров на портативные электронные устройства [53].

Наиболее похожей на SentiStrength программой является SO-CAL, которая, однако, была создана для других целей. Она использует лексический метод для классификации текстов на положительные или отрицательные, а также использует словарь слов, которым присвоена оценка по общей для двух тональностей шкале от -5 до +5 [45]. Словарь в SO-CAL был создан на основе корпуса, размеченного людьми, а также с использованием словаря General Inquirer. В итоге получилось 2,252 прилагательных, 745 наречий 1,142 существительных и 903 глаголов, для всех существительных и глаголов была проведена лемматизация. Каждому слову

[2] Affective Norms for English Words

[3] Linguistic Inquiry and Word Count (лингвистическое исследование и подсчет слов) — программное обеспечение для анализа текстов, для вычисления частотности использования слов человеком в электронных письмах, в речи, стихах и др. Позволяет определить степень эмоциональной нагрузки текста.

было присвоено значение «приоритетной тональности» — стандартной тональности данного слова на всём множестве контекстов, а не в отдельно взятом. В SO-CAL также имеется список их как минимум 187 эмоциональных выражений, состоящих из нескольких слов, а также список усилительных выражений, которые повышают или понижают силу тональности и алгоритм, который справляется с отрицанием. Итоговая тональность предложения складывается из среднего арифметического значения тональности всех слов, найденных в нём после всех преобразований. Данная программа показывает хорошие результаты для сбалансированного набора данных, состоящих в основном из текстов с различных сайтов и новостных текстов [45].

3.3. Источник данных для создания SentiStrength

Источником тестовых данных для разработки SentiStrength была выбрана социальная сеть MySpace, потому что это общественная среда, содержащая огромное количество текстов на ненормативном языке, которая, по состоянию на 2009 год, стала одним из самых посещаемых веб-сайтов в мире. Из MySpace была взята случайная выборка из 40000 комментариев, принадлежащих каждому 15-му пользователю, зарегистрированному 18 июня 2007 года. Далее были выбраны комментарии людей, в профиле которых утверждалось, что они граждане США и не являются музыкантами, комиками или кинорежиссёрами. Из них игнорировались две группы людей: к первой принадлежали люди, у которых было менее, чем 2 друга (они рассматривались как неактивные), ко второй — люди с более чем 1000 друзей или 4000 комментариев (для них прослеживалась необычно большая активность). Далее для каждого оставшегося пользователя был определён оставивший комментарий друг, который удовлетворял тем же критериям, что описаны выше. В результате было выбрано по одному случайному комментарию от друга к пользователю и наоборот. Комментарии были извлечены в декабре 2008г. В результате получилась огромная случайная выборка сообщений граждан

США. В последствии были удалены сообщения, содержащие спам, письма, рассылаемые по цепочке нескольким адресатам, и комментарии, содержащие картинки.

Хотя анализ тональности обычно связан с мнениями [38], Уилсон в своей работе [56] распространил его также на психологическую задачу определения из текста скрытого внутреннего состояния автора. Однако для данных из MySpace задача была не определить мнения или внутреннее состояние автора, а определить роль выраженной тональности для общения онлайн. Отсюда, исследование было сосредоточено на выявлении тональности, выраженной в каждом сообщении, вне зависимости от того, отражает ли она скрытое внутреннее состояние автора, интерпретацию подразумеваемого сообщения или скрытое внутреннее состояние читателя.

Для того, чтобы получить надёжную оценку экспертов относительно случайной выборки комментариев из MySpace, были проведены два пробных аннотирования с использованием двух разных наборов данных (общим объёмом в 2600 комментариев). Аннотирование было проведено с целью выявления подходящей шкалы и главных проблем при оценке. Для экспертов была составлена и затем усовершенствована инструкция, а также была создана онлайн-система для случайного выбора комментариев, которые были представлены экспертам. Одним из главных результатов пробной разметки явилось то, что эксперты относили восклицательные предложения к положительным, если в них не был явно выражен отрицательный контекст. Например, фраза «Привет!!!» интерпретировалась большинством экспертов как положительная по умолчанию, так как она выражает интенсивность в контексте, который не даёт никаких подсказок относительно тональности эмоции. И наоборот, фразу «Неудачник!!!» эксперты оценили как более отрицательную, чем просто «Неудачник», так как знаки восклицания ассоциируются с отрицательным словом. В последствии, инструкции были пересмотрены, и в них было явно занесено утверждение, что объединение, по-

видимому, нейтрального восклицания и положительной тональности является допустимым.

Для окончательного решения было выбрано более тысячи комментариев из MySpace (в среднем с 20 словами и 101 символами на комментарий) для их оценки по следующей пятибалльной шкале для положительной и отрицательной тональности:

(нет положительной эмоции или интенсивности) **1– 2 – 3 – 4 – 5** (крайне положительная эмоция)

(нет отрицательной эмоции) **1– 2 – 3 – 4 – 5** (крайне отрицательная эмоция)

Экспертам были предоставлены как устные инструкции для оценки текстов, так и брошюра, объясняющая задание (составленная на основе работы [54]). Самые значимые инструкции приводятся в четвёртой главе данной работы. Начальная версия брошюры содержала примеры комментариев с соответствующей положительной или отрицательной оценкой, но на практике данная информация никак не помогала экспертам во время пробной оценки. По этой причине набор комментариев не использовался, так что степень согласия между экспертами могла быть оценена более реалистично, так как исключалась возможность того, что некоторые комментарии были слишком похожи на те, что встречались в задании.

Разными людьми эмотиконы воспринимались по-разному отчасти из-за их жизненного опыта, а также из-за личных разногласий и даже пола автора. Для разработки системы оценка должна давать последовательный взгляд на тональность в данных, а не среднестатистическое восприятие людей. Как результат, были привлечены эксперты одного пола (женского), пробная оценка проводилось для того, чтобы отобрать людей, которые давали однородные результаты. Изначально было выбрано 5 экспертов, однако результаты двоих из них впоследствии оказались непригодными из-за аномальности: один эксперт давал значительно более высокие оценки комментариям с положительной тональностью, другой давал в целом непоследовательные результаты. Для троих экспертов были вычислены и округлены средние значения по каждому

комментарию. Из этого получился золотой стандарт для экспериментов. Ниже приведено несколько примеров текстов и оценок:

- hey witch wat cha been up too (оценки: +: 2,3,1; -: 2,2,2)
- omg my son has the same b-day as you lol (оценки: +: 4,3,1; -: 1,1,1)
- HEY U HAVE TWO FRIENDS!! (оценки: +: 2,3,2; -: 1,1,1)
- What's up with that boy Carson? (оценки: +: 1,1,1; -: 3,2,1)

Разработчиками была посчитана степень согласия между тремя экспертами. Для этого был использован коэффициент, известный под названием альфа Криппендорфа [27]. Значение альфы для предложений с положительной тональностью получилось 0,5743, для отрицательных — 0,5634. Данные результаты являются довольно надёжными, это показывает, что оценки экспертов во многих случаях совпадают. Однако, они являются недостаточно надёжными, так как нормой считаются значения <0,67. Тем не менее, использование средней оценки экспертов в качестве золотого стандарта казалось вполне разумным методом получения оценок тональности.

3.4. Описание алгоритма SentiStrength

Система SentiStrength автоматически производит анализ до 16.000 текстов из социальных сетей в секунду с точностью, сравнимой с точностью человека, для английского языка.

Программа оценивает силу тональности для каждого сообщения по двум шкалам одновременно:

-1 (не отрицательный) до -5 (крайне отрицательный)

1 (не положительный) до 5 (крайне положительный)

Две шкалы используются потому, что исследования из области психологии выявили, что мы обрабатываем положительные и отрицательные мнения параллельно (например, [19]). Отсюда — смешанные чувства и эмоции.

SentiStrength также может выдавать результаты на основе бинарной классификации (положительный/отрицательный), на основании классификации с тремя классами (положительный/отрицательный/ нейтральный). Также она может оценивать мнение по одной шкале (-4; 4).

Изначально программа создавалась для анализа коротких текстов на английском языке, но она также может быть настроена на другие языки путём изменения файлов исходных данных.

Данная программа находится в свободном доступе, её можно свободно использовать для научных исследований.

Учитывая то, что программа основана на использовании словаря эмоциональных слов, её можно настроить также и на определённую предметную область.

Описание алгоритма программы [49]:

- Основа алгоритма — это список эмоциональных слов. Каждому слову в списке присваивается значение интенсивности выражаемой им эмоции от 2 до 5. Как и в словаре LIWC, в некоторых словах изменяемая часть заменена на символ-джокер звёздочку (*), например, для слова «ador*» в тексте будут находиться словоформы «adorable», «adored», «adoring» и т.д. Единственным словом в словаре, имеющим одновременно положительную и отрицательную оценку 2, является слово «miss» в значении скучать по кому-то (оно часто используется для выражения грусти и любви одновременно).
- В анализ встроен обучающий алгоритм для того, чтобы оптимизировать значения эмоций в словаре, размеченном человеком.
- Имеется также алгоритм исправления правописания. Программа автоматически удаляет буквы, повторяющиеся более двух раз (пример — helllo -> hello), удаляет удвоенную букву, если такое сочетание в английском языке встречается редко (пример — niice -> nice), удаляет удвоенную букву в слове с ошибкой, если после исправления

получится нормативное слово без ошибки (например, исправится nnice -> nice, но не hoop -> hop и не baaz -> baz).

- Программа использует список усилительных слов, которые могут увеличивать или уменьшать значение тональности. Значение каждого слова повышается на 1 или 2 единицы (например, с наречиями «very», «extremely») или уменьшается на 1 единицу (со словом «some»).
- Используется список слов, выражающих отрицание. Они меняют значение тональности на противоположное. Распознавание случаев, когда присутствие отрицательных слов не меняет тональность, не было встроено, так как такие случаи встречались редко в пробном наборе данных.
- Используется список со эмотиконами. Так как эмотиконы являются более независимыми от предметной области, чем слова, их можно использовать как относительно надёжный показатель тональности в предложениях.
- Если в предложении встречаются эмоциональные слова, которые требуют исправления (с повторяющейся более двух раз буквой), у таких слов значение тональности повышается на 1 единицу. В сообщениях из социальной сети MySpace это считается обычным средством выражения эмоций и силы эмоций. Однако одна лишняя буква считается опечаткой.
- Если предложение оканчивается на восклицательный знак, значение тональности повышается на 2 единицы.
- В вопросительных предложениях игнорируются отрицательные слова. Например, предложение “are you angry?” будет классифицировано как нейтральное.

Описанная выше версия SentiStrength показывает точность в 60,6% для положительной тональности, и в 72,8% для отрицательной. В первом случае программа оценивает тональность существенно лучше, чем другие инструменты

(лучшим результатом оценки положительных предложений на момент исследования была точность в 58.5%).

В последующих версиях программы [48; 47] были предприняты попытки улучшить эффективность работы программы для определения отрицательной тональности.

3.5 Выводы к главе 3

В этой главе мы рассмотрели основные предпосылки создания программы SentiStrength и место алгоритма SentiStrength в системе других методов сентимент-анализа. Инструмент был разработан Майком Телваллом, Кеваном Бакли и их коллегами в университете Уольверхэмптон в 2010 году. Программа оценивает силу тональности коротких сообщений одновременно по двум шкалам (положительной и отрицательной) от 1 до 5 и от -1 до -5. Данная система основана на использовании словаря эмоциональной лексики и корректирующих правил. Программа разрабатывалась на основе сообщений из социальной сети MySpace. Словарь был частично дополнен лексикой из словаря LIWC. В главе подробно описан процесс создания и алгоритм работы SentiStrength.

Алгоритм данного инструмента эффективнее других методов справляется с анализом положительной тональности коротких неформальных текстов. Результат работы программы оценивался с помощью точности и коэффициента корреляции между оценками экспертов и оценками программы. Позднее разработчики предприняли несколько попыток улучшения работы программы для отрицательной тональности путём расширения исходных данных.

Глава 4. Настройка системы SentiStrength на украинский язык

4.1. Обзор предыдущих работ по анализу тональности текстов на украинском языке

Нам удалось найти лишь два исследования на тему анализа тональности текстов на украинском языке. В статьях [6; 32] описаны этапы создания тонального словаря с использованием аннотированного корпуса текстов. Полный словарь, однако, нигде не представлен и, кроме того, он был составлен для анализа конкретной предметной области — ресторанных отзывов.

Кроме того, анализом тональности текстов на украинском языке занимается исследовательская группа lang-uk [60]. Lang-uk — это сообщество специалистов в области компьютерной обработки текстов (программистов, лингвистов, исследователей), основными направлениями работы которого являются:

- сбор и публикация корпусов и других наборов текстовых данных на украинском языке;
- создание моделей на основе корпусов для решения прикладных задач обработки украинских текстов;
- имплементация этих моделей в ряде публично-доступных микросервисов.

В рамках одного из проектов группа разработала общий тональный словарь для украинского языка, содержащий 3442 слов, имеющих не нейтральную тональность (-2, -1, 1, 2). Для создания словаря было использовано два метода — ручной (средняя оценка нескольких экспертов) и автоматический (с помощью моделей word2vec и lex2vec). В словаре все слова по возможности приведены к начальной грамматической форме, а наречия заменены на однокоренные прилагательные. Словарь находится в открытом доступе. На одном из этапов настройки SentiStrength этот словарь был использован для расширения исходных данных.

4.2. Файлы исходных данных системы SentiStrength

Для настройки SentiStrength версии 2.2 на украинский язык для начала необходимо было изменить восемь файлов исходных данных. В их число входит список усилительных слов, список с эмотиконами, список эмоциональных слов, идиом, иронических, отрицательных, вопросительных слов и список сленговых выражений. Ключевыми для работы системы являются описанные ниже файлы.

EmotionLookupTable.txt — этот файл содержит список тональных слов, таких как «любить», «ненавидеть». Каждое слово должно иметь оценку, которая указывает на то, какую типичную тональность, а также какую типичную силу тональности оно выражает в соответствии со следующей шкалой:

-5 Очень сильная отрицательная тональность (напр., мучительно)

-4 Сильная отрицательная тональность (напр., ненавидеть)

-3 Умеренно отрицательная тональность (напр., неприязнь)

-2 Слегка отрицательная тональность (напр., неудобство)

2 Слегка положительная тональность (напр., удовлетворять)

3 Умеренно положительная тональность (напр., счастливый)

4 Сильная положительная тональность (напр., влюблённый)

5 Очень сильная положительная тональность (напр., восторженный)

Стоит отметить, что в этом и последующих словарях слова должны быть употребительны в неформальной письменной речи.

Оценка 1 (что означает нулевую тональность) может быть присвоена словам, которые, по мнению составителя словаря, не выражают тональность, но могут выражать в некоторых контекстах. Это означает, что составитель принимает во внимание данное понятие.

Необходимо поместить словарь в текстовый файл, по одному слову на строку, за словом через tab следует одна из оценок -5, -4, -3, -2, 2, 3, 4, 5. Для правильной работы программы все файлы, которые содержат символы, отсутствующие в английском языке, необходимо сохранять в кодировке UTF-8.

Если у слова есть несколько окончаний, которые не меняют его тональность, то в этом случае окончание заменяется на символ-джокер звёздочку (*). Например, такому варианту слова «hate*» будут соответствовать слова hate, hater и hated, и, таким образом, не нужно создавать отдельные словарные статьи для этих слов. Однако символ-джокер не рекомендуется ставить, если из-за него будут совпадать несвязанные по смыслу слова. Например, слову «amaz*» будут соответствовать amazing, amazed (с положительной тональностью) и Amazon (нейтральное слово), что является неверным.

Разработчикам во второй версии программы SentiStrengh частично удалось решить эту проблему. Для более длинных нейтральных слов, которые совпадают с тональными (как в случае с amazed и Amazon), они создавали отдельную словарную статью с весом 1 или -1.

BoosterWordList.txt — это файл, содержащий список слов («очень», «немного»), которые повышают или понижают тональность следующих за ними слов. Каждому слову присваивается значение, насколько оно усиливает или ослабляет тональность, в соответствии со следующей шкалой:

-2 Сильное понижение тональности (напр., плохо)

-1 Небольшое понижение тональности (напр., немного)

1 Небольшое повышение тональности (напр., очень)

2 Сильное повышение тональности (напр., чрезвычайно)

Оформляется словарь таким же образом, как и предыдущий.

NegatingWordList.txt — этот файл содержит список слов, которые практически всегда будут указывать на то, что смысл предложения, слова или словосочетания отрицается («не», «никогда»).

QuestionWords.txt — файл, в котором содержится список слов, которые практически всегда указывают на то, что предложение является вопросительным («как», «когда», «почему»).

4.3. Создание словарей для украинского языка

При создании украинской версии SentiStrengh мы опирались на существующую версию для русского языка, так как два этих языка являются родственными.

Первой задачей было создание словарей для украинского языка. Вначале был осуществлен перевод русских словарей на украинский язык. Перевод был выполнен автоматически, с помощью веб-службы Яндекс.Переводчик. Правильность автоматического перевода проверялся двумя экспертами, носителями украинского языка, и, по возможности, пополнялся синонимами. В результате получился список из 783 единиц, в которые входят 63 усилительных слова, 506 отрицательных, 29 потенциально тональных и нейтральных (с оценкой 1 и -1), 172 положительных слова, 6 слов, меняющих тональность на противоположную, и 7 вопросительных слов. Была учтена ненормативная лексика и сленговые выражения.

В тональном словаре в основном представлены основы слова. Было принято решение не использовать алгоритмы стемминга, потому что словарь по большей части содержит ненормативную лексику, так как направлен для обработки неформальных текстов из социальных сетей. В данном случае обычные стеммеры не справились бы со своей задачей.

Приведем несколько примеров для каждого словаря:

Словарь усилительных слов (BoosterWordList):

абсолютно 2
дуже 2
злегка 1
настільки 1
несказанно 3
особливо 1
просто 1
трохи -1

Эмоциональная лексика (EmotionLookupTable):

агрес* -4

агресивн* -3
азарт* 3
акуратн* 2
ангельськ* 3
апетитн* 3
афіген* 4
беззакон* -2
безнаді* -3
безпереч* 1
безперіч* 1
мімішн* 3
мінорн* -2
міф* -2

Вопросительные слова (QuestionWords):

де
коли
навіщо
хто
чому
що

Слова, меняющие тональность на противоположную (NegatingWordList):

без
не
нема
немає
ні
ніколи

4.4. Создание золотого стандарта и обучение программы

Вторым этапом работы стало создание золотого стандарта для обучения программы и проверка результатов с помощью тестовой выборки.

Для создания золотого стандарта необходимо было собрать 1200 твитов на украинском языке. Это было сделано с помощью программы Webometric analyst. Для поиска украиноязычных записей были использованы самые частотные слова украинского языка («що», «від», «є», «які», «було», «чи», «вже», «якщо», «щоб», и

т.д), взятые из электронного частотного словаря [67]. Из полученной выборки были удалены повторяющиеся твиты, а также тексты на русском и белорусском языках. В итоге получилось 1954 текста, из которых случайным образом было выбрано 1000 твитов для золотого стандарта и 200 твитов для тестовой выборки.

Для оценки твитов были привлечены 3 эксперта, 3 носителя украинского языка. При оценке экспертам нужно было придерживаться следующих инструкций, основу для которых составили инструкции из [54]:

- необходимо оценивать каждый короткий текст как содержащий и положительную, и отрицательную тональность, причём одна не должна отменять другую («я тебя люблю и ненавижу);
- не рекомендуется советоваться и обсуждать с кем-либо решения;
- формальных критериев для оценки не существует, всё, что нужно — это использовать свои знания и интуицию;
- необходимо рассматривать каждый отдельный текст независимо от остальных, но стараться быть последовательным в своих решениях;
- необходимо интерпретировать эмоцию внутри отдельного текста в том виде, в каком она выражена, и игнорировать все остальные тексты;
- необходимо также оценивать тексты с точки зрения их содержания, а не с точки зрения эмоционального состояния автора или предполагаемого состояния читателя

Для удобства работа проводилась в программе Microsoft Excel, у каждого эксперта в листе было также поле «Комментарии». Это поле необходимо было использовать только в том случае, если, по мнению экспертов, с данными было что-то не так или если они хотели привлечь к чему-то наше внимание. Например, необходимо было отметить, если твит написан полностью на другом языке, если в нём вообще нет текста или если случайно содержит информацию, связанную со сбором данных. Эти комментарии помогли привести выборку в порядок.

После того, как твиты были оценены, для золотого стандарта был посчитан уровень согласия между тремя экспертами и для каждой пары по отдельности. Для этого также был использован коэффициент альфа Криппендорфа. Коэффициент был посчитан при помощи онлайн утилиты ReCal [62], разработанной в 2008 году Дином Фрилоном, аспирантом Вашингтонского университета. Для троих экспертов альфы получились 0,581 и 0,542, а средний процент согласия 77% и 73% для положительных и отрицательных оценок соответственно. Результаты, как и у разработчиков, получились удовлетворительные, однако, нормой считаются значения 0,67 и выше. Ниже в таблице 1 приведены результаты для каждой пары экспертов. Представлены коэффициенты согласия для точных совпадений и для совпадений с разницей в единицу (оценки 2 и 3, 3 и 4, 4 и 5 считались за согласие), был также подсчитан % согласия.

Таблица 1

	Положительные предложения				Отрицательные предложения			
	точное совп.	% согл.	±1	% согл.	точное совп.	% согл.	±1	% согл.
Эксперт 1 и эксперт 2	0,701	84,1	0,819	90,4	0,74	83,9	0,858	91,2
Эксперт 1 и эксперт 3	0,646	80,3	0,766	87	0,524	73,1	0,646	80,1
Эксперт 2 и эксперт 3	0,397	66,8	0,858	91,2	0,344	62,6	0,517	72,4

Уровень согласия между каждой парой экспертов

Из таблицы видно, что коэффициент согласия между экспертом 2 и 3 весьма низкий. Это значит, что у эксперта 3 несколько другие представления об эмоциях, особенно отрицательных. Так как общее согласие между тремя людьми удовлетворительное, было принято решение не менять экспертов. Однако в

дальнейшем при создании других золотых стандартов, с экспертом, оценки которого существенно отличаются от остальных, рекомендуется обсудить несогласия или, возможно, даже заменить его.

Далее для каждого текста из выборки было посчитано и округлено до целых среднее арифметическое оценок трёх экспертов. Таким образом, было завершено создание золотого стандарта и тестовой выборки.

После того, как все файлы со словарями были помещены в нужную директорию (C:\SentStrength_Data), была проведена оценка работы программы.

Тексты из тестовой выборки были помещены в текстовый файл по одному твиту на строку в следующем формате, чтобы программа автоматически посчитала процент точных совпадений полученных оценок с оценками экспертов:

[положительная оценка] tab [отрицательная оценка] tab [текст]

В пункте меню программы «Sentiment Strength Analysis» была выбрана опция «Analyse ALL Texts in File» для того, чтобы программа оценила тексты в нашем файле.

В файле с результатами к имеющимся данным в исходном файле через tab была добавлена следующая информация: дублируется текст с тегом
, который маркирует конец предложения после знака препинания; далее следуют положительная и отрицательная оценки программы, а затем строка такого типа:

Це[0] офігенне[4] відчуття[0] коли[0] у[0] тебе[0] є[0] божевільні[0] друзі[0] [[Sentence=-1,5=word max, 1-5]] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/ujbyWoMml0[0] [[Sentence=-1,1=word max, 1-5]][[[5,-1 max of sentences]]]

Цифры в квадратных скобках означают силу тональности, которое имеет каждое слово. В силу особенностей работы алгоритма, эти веса отличаются от тех, что закодированы в словаре, на 1 (слово «офіген*» в словаре имеет оценку 5, здесь в предложении — оценку 4). Каждому предложению по отдельности присваиваются значения тональности (оценки слов в предложении суммируются, если сумма больше 5, то она приравнивается к 5) — «Sentence=-1,5». В самом конце тексту присваиваются максимальные значения тональностей среди оценённых

предложений — «5,-1 max of sentences». Таким образом, предложение «Це офігенне відчуття, коли у тебе є божевільні друзі. https://t.co/ujbyWoMml0» получило оценку 5, -1. Некоторые другие примеры оценённых программой твитов представлены в приложении 2.

Для данного файла с результатами была посчитана точность работы программы. Результаты представлены в таблице 2.

Таблица 2

Положительная тональность		Отрицательная тональность	
точное совпадение	совпадение ±1	точное совпадение	совпадение ±1
63%	71%	60%	68%

Точность работы программы после создания тонального словаря

На основании оценки результатов можно сделать вывод, что программа справилась с анализом достаточно хорошо. Ошибки можно разделить на следующие группы:

- Случаи, когда в нейтральном предложении используются тональные слова (программа относила документы в этих случаях либо к положительным, либо к отрицательным отзывам).

«З чим воюють українські бійці: від «ретро» до саморобних новинок (Відео) #news»

Текст новостной, нейтральный, однако из-за слова «воювати» (воевать) программа отнесла данный твит к отрицательным и присвоила ему оценки 1 и -3.

- Случаи, когда в тексте тональность присутствует, но выражена неявно.

«вирішили з ма, що купуємо мені гітару і я вчитимусь грати, а ще видамо книгу наших старих віршів. 01:20 відверті розмови в кухні під коньяк»

Эксперты отнесли этот твит к положительным, а программа к нейтральным.

- Случаи, когда тональность определялась правильно, но значение, присвоенное программой, не совпадало со значением, присвоенным носителем языка.

«рятуйте! Як відшити хорошого друга який підкатує, але так, щоб не образити? (розповіла що стосунків не хочу, але він не відчепився)»

Эксперты присвоили данному тексту оценки 2 и -3, а программа 4 и -2.

Данные типы ошибок были проанализированы и выявлены причины их возникновения:

1) Лакуны в тональном словаре.

2) В некоторых случаях возникала проблема с символом-джокером, и более длинные нейтральные слова оценивались как положительные или отрицательные.

3) Омонимия и многозначность слов. Например, слово «лютий» в украинском языке является названием месяца февраля, а также является прилагательным со значением «злой, лютый». Также слово «мат», как и в русском языке имеет несколько значений: является синонимом слова «матовость» (когда речь идёт о поверхности), слова «ругань» или является шахматным термином.

4) Веса в словарях не были оптимизированы с помощью золотого стандарта.

5) Слишком подробная система оценки (четыре уровня для каждой тональности).

Первую причину можно устранить путём простого пополнения словаря. Вторую — создав отдельные словарные статьи для коротких эмоциональных слов и их словоформ. Третью причину можно лишь частично устранить, анализируя частотность появления слов в конкретном значении в коротких неформальных текстах. Четвертая причина устраняется путём использования обучающего алгоритма, встроенного в программу. Пятая причина может быть исправлена с помощью уменьшения числа уровней для каждой тональности или с помощью отдельного модуля, который нивелировал бы различия между близкими уровнями оценки.

Была проведена попытка улучшить словарь с учётом описанных выше проблем. Получился в итоге словарь объёмом 1205 слов, в которые входят 70 усилительных слов, 705 отрицательных, 42 потенциально тональных и

нейтральных (с оценкой 1 и -1), 375 положительных слова, 6 слов, меняющих тональность на противоположную, и 7 вопросительных слов.

Для некоторых коротких эмоциональных слов («біс», «вина», «гарний» и др.) вместо использования символа-джокера вводилась целая парадигма, чтобы избежать совпадения с нейтральными словами. В некоторых случаях вводились основы нейтральных слов. Также на этом этапе исходные данные были частично дополнены за счёт тонального словаря, разработанного группой lang-uk.

Полученный словарь проверялся двумя способами. Была проведена перекрёстная проверка с использованием золотого стандарта. В основе данной проверки лежало разделение исходного множества данных (1000 текстов) на 9 частей. В настройках программы есть возможность самостоятельно задавать число частей. Чтобы провести перекрёстную проверку, в пункте меню программы «Sentiment Strength Analysis» необходимо выбрать опцию «Run 10-fold cross-validation to assess above optimisation algorithm [n times]».

Результаты перекрёстной проверки представлены в таблице 3.

Таблица 3

Положительная тональность			Отрицательная тональность		
Corr+	Acc+	AccWithin1+	Corr-	Acc-	AccWithin1-
0,4995	61,06%	87,69%	0,485	59,36%	87,29%

Результаты перекрёстной проверки обучающего алгоритма

В таблице 3 Corr+ и Corr- означают корреляцию между оценками экспертов и оценками программы для положительной и отрицательной тональности соответственно. Значения выше 0,2 являются хорошим результатом, это означает, что программа работает, однако её работу можно улучшить путём расширения словаря. Значения выше 0,4 являются превосходным результатом (для сравнения, версия для английского языка показывает результаты в диапазоне 0,45-0,55).

Acc+ и Acc- показывают точность работы программы относительно присвоенных экспертами положительных и отрицательных оценок соответственно.

AccWithin1+ и AccWithin1- показывают точность работы программы с разницей в ±1 относительно присвоенных экспертами положительных и отрицательных оценок соответственно.

Далее было решено использовать золотой стандарт для получения словаря с оптимальными весами.

В словаре произошло 23 изменения, некоторые из них:

+1 проти
+1 улюблен*
+1 хорош*
+1 дяку*
-1 ненави*
+1 крик*
-1 допомог*
-1 коха*
-1 представ*
+1 помер*

Цифры перед словами означают, на сколько единиц увеличился или уменьшился вес слова в словаре.

Новый словарь был помещён в уже известную нам директорию — C:\SentStrength_Data — для того, чтобы оценить его эффективность с помощью тестовой выборки. Полученные результаты представлены в таблице 4.

Таблица 4

Положительная тональность		Отрицательная тональность	
точное совпадение	совпадение ±1	точное совпадение	совпадение ±1
73%	82%	70%	78%

Точность работы программы после оптимизации весов в словаре

На основании результатов проверки алгоритма можно сделать вывод, что программа работает хорошо для нового языка. Количество ошибок сократилось как минимум на 10% для каждого случая. Тем не менее, при анализе последних результатов прослеживаются все типы ошибок, описанные выше. Стоит также отметить, что программа работает лучше для положительной тональности.

Случаи с новостными нейтральными текстами, содержащими эмоциональные слова, предположительно можно решить следующим образом: предварительно провести классификацию текстов на новостные и не новостные, а затем работать только со второй группой сообщений.

Для улучшения работы программы в дальнейшем необходимо осуществить следующие этапы и, при желании, повторить процесс несколько раз:

- создать новый золотой стандарт объёмом как минимум 1000 текстов;
- расширить исходные данные;
- оптимизировать веса в словаре;
- оценить работу алгоритма.

По возможности необходимо также создать словарь эмоциональных идиом украинского языка, а также словарь ироничных высказываний.

Логичным шагом будет также сравнение работы данной программы с другими алгоритмами анализа тональности для украиноязычных текстов.

4.5 Выводы к главе 4

В этой главе описывается эксперимент по настройке системы SentiStrength на украинский язык. Настройка включала в себя следующие этапы:

1) создание тонального словаря;
2) создание золотого стандарта и тестовой выборки;
3) проверка словаря на тестовой выборке и анализ результатов;
4) улучшение и расширение исходных данных;
5) оптимизация весов с помощью встроенного обучающего алгоритма;
6) проверка улучшенного словаря на тестовой выборке и подведение итогов.

Оценка результатов на разных этапах проводилась с помощью тестовой и обучающей выборки объемом 200 и 1000 сообщений соответственно из социальной сети твиттер. Были показаны следующие результаты:

После первой проверки на тестовой выборке точность при полном совпадении оценок алгоритма с оценками экспертов составляла 63% и 60% для положительных и отрицательных сообщений соответственно, а точность с разницей в единицу (оценки 2 и 3, 3 и 4, 4 и 5 считались за совпадение) получилась 71% и 68%. После второй проверки результаты удалось улучшить как минимум на 10% в каждом случае.

Перекрёстная проверка на обучающей выборке показала также хорошие результаты: коэффициенты корреляции между оценками программы и оценки экспертов составили 0,4995 и 0,485 для положительных и отрицательных сообщений соответственно.

Заключение

Анализ тональности сообщений — это быстро развивающаяся область компьютерной лингвистики, открывающая большие возможности для различных лингвистических, социологических, психологических исследований и перспективы в коммерческом применении.

Программа SentiStrength является эффективным инструментом для оценки силы тональности коротких сообщений, написанных на неформальном языке. Главной причиной достаточно успешной работы алгоритма является возможность распознавания слов с ненормативной орфографией.

Результаты данной работы показали, что анализ сообщений из социальной сети твиттер является довольно непростой задачей в силу изобретательности пользователей в языковых выражениях, передачи тональности без использования эмотивной лексики и различных взглядов на тональность экспертов, кодирующих данные. Последнее означает, что, по-видимому, не существует истинно верной классификации для многих сообщений.

В рамках данной работы были достигнуты следующие результаты:

1) была изучена предметная область анализа тональности текста, было разобрано применение анализа, его задачи и основные методы;
2) был детально описан процесс работы систем программы SentiStrength;
3) был проведен процесс настройки данной системы на украинский язык, улучшение работы алгоритма было достигнуто с применением методов машинного обучения;
4) была проведена оценка результатов работы созданной системы.

При последней оценке работы программы были получены следующие значения точности: 73% и 70% для положительных и отрицательных сообщений соответственно при однозначном совпадении и 82% и 78% для совпадений с разницей в единицу.

Результаты работы были представлены на студенческой конференции филологического факультета СПбГУ в апреле 2017 года.

Полученные данные будут добавлены в систему SentiStrength университета Уольверхэмптон в Великобритании со ссылкой на кафедру математической лингвистики СПбГУ.

Список литературы

1. *Клековкина М.В., Котельников Е.В.* Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики (рус.). RCDL-2012, Переславль-Залесский, Россия: конференция, 2012.
2. *Котельников Е.В., Клековкина М.В.* Автоматический анализ тональности текстов на основе методов машинного обучения.
3. *Пазельская А., Соловьев А.* Метод определения эмоций в текстах на русском языке. The international conference on computational linguistics and intellectual technologies “Dialogue 2011”: конференция. Москва, 2011. с. 510-522. Что такое тональность.
4. *Паничева П.* Система сентиментного анализа ATEX, основанная на правилах, при обработке текстов различных тематик. Sentiment Analysis Track at ROMIP, 2012.
5. *Поляков П.Ю., Калинина М.В., Плешко В.В.* Исследование применимости методов тематической классификации в задаче классификации отзывов о книгах. ООО «ЭР СИ О», Москва, Россия.
6. *Романишин М., Романюк А.* Тональний словник української мови на основі сентимент-анотованого корпусу. Українське мовознавство, 2013. Вип. 43, с. 63-74.
7. *Asur Sitaram and Bernardo A. Huberman.* Predicting the future with social media. Arxiv preprint arXiv: 1003.5699, 2010.
8. *Babbar Rohit, Partalas Ioannis, Gaussier Eric, Amini Massih-Reza*. On Flat versus Hierarchical Classification in Large-Scale Taxonomies.
9. *Baccianella, S., Esuli, A., & Sebastiani, F.* (2010). SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. Proceedings of the Seventh conference on International Language Resources and Evaluation, pp. 2200-2204.
10. *Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., Goot, E. v. d., Halkia, M., Pouliquen, B., & Belyaeva, J.* (2010). Sentiment analysis in the news. In Proceedings

of the international conference on language, resources and evaluation, pp. 2216-2220. Valletta, Malta.

11. *boyd, d.* (2008). Taken out of context: American teen sociality in networked publics. University of California, Berkeley, Berkeley.
12. *boyd, d.* (2008). Why youth (heart) social network sites: The role of networked publics in teenage social life. In D. Buckingham (Ed.), Youth, identity, and digital media, pp. 119-142. Cambridge, MA: MIT Press.
13. *Bradley, M. M., & Lang, P. J.* (1999). Affective Norms for English Words (ANEW): Stimuli, instruction manual, and affective ratings (Tech. Report C-1). Gainesville: University of Florida, Center for Research in Psychophysiology.
14. *Brill, E.* (1992). A simple rule-based part of speech tagger. Proceedings of the Third Conference on Applied Natural Language Processing, pp. 152-155.
15. *Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, K.P.* Measuring User Influence in Twitter: The Million Follower Fallacy. Proceedings of the 4th International AAAI Conference on Weblogs and Social Media (ICWSM), Washington, May 2010.
16. *Choi, Y., & Cardie, C.* (2008). Learning with compositional semantics as structural inference for subsentential sentiment analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 793-801.
17. *Das, S., & Chen, M.* (2001). Yahoo! for Amazon: Extracting market sentiment from stock message boards. Proceedings of the Asia Pacific Finance Association Annual Conference (APFA), Bangkok, Thailand, July 22-25, Дата доступа 28 апреля 2017г из: http://sentiment.technicalanalysis.org.uk/DaCh.pdf.
18. *Derks, D., Bos, A. E. R., & von Grumbkow, J.* (2008). Emoticons and online message interpretation. Social Science Computer Review, 26(3), pp. 379-388.
19. *Fox, E.* (2008). Emotion science. Basingstoke: Palgrave Macmillan, p. 127.
20. *Freitas A.A., de Carvalho A.C.P.L.F.* (2007) Research and trends in data mining technologies and applications: tutorial on hierarchical classification with applications in bioinformatics.

21. *Fullwood, C., & Martino, O. I.* (2007). Emoticons and impression formation. The Visual in Popular Culture, 19(7), pp. 4-14.
22. *Gamon, M., Aue, A., Corston-Oliver, S., & Ringger, E.* (2005). Pulse: Mining customer opinions from free text (IDA 2005). Lecture Notes in Computer Science, 3646, pp. 121-132.
23. *Ghazi Diman, Inkpen Diana, Szpakowicz Stan.* Hierarchical versus Flat Classification of Emotions in Text. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text, pp. 140-146, Los Angeles, California, June 2010.
24. *Joshi Mahesh, Dipanjan Das, Kevin Gimpel, and Noah A. Smith.* Movie reviews and revenues: An experiment in text regression. In Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference (NAACL 2010), 2010.
25. *Jurafsky Daniel, Martin James H.* Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition. Pearson Education International, 2009. 1024 pp.
26. *Kan D.* Rule-based approach to sentiment analysis. Sentiment Analysis Track at ROMIP, 2011.
27. *Krippendorff, K.* (2004). Content analysis: An introduction to its methodology. Thousand Oaks, CA: Sage.
28. *Kukich, K.* (1992). Techniques for automatically correcting words in text. ACM computing surveys, 24(4), pp. 377-439.
29. *Liu Bing.* Sentiment Analysis and Opinion Mining. Morgan & Claypool Publishers, May 2012.
30. *Liu Bing*. Sentiment Analysis Tutorial. AAAI-2011, San Francisco, USA.
31. *Liu Yang, Huang Xiangji, An Aijun, Yu Xiaohui*: ARSA: a sentiment-aware model for predicting sales performance using blogs. SIGIR 2007: pp. 607-614.
32. *Lobur M., Romaniuk A., Romanyshyn M.* Defining an approach for deep sentiment analysis of reviews in Ukrainian. Вісник Національного університету "Львівська

політехніка". Комп'ютерні системи проектування. Теорія і практика, 2012. № 747, с.124-130.

33. *Martineau Justin, and Finin Tim*. Delta TFIDF: An Improved Feature Space for Sentiment Analysis. Third AAAI Internatonal Conference on Weblogs and Social Media, May 2009, San Jose CA.
34. *Neviarouskaya A., Prendinger H., & Ishizuka M.* (2007). Textual affect sensing for sociable and expressive online communication. Lecture Notes in Computer Science, 4738, pp. 218-229.
35. *O'Connor Brendan, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith*. From Tweets to Polls: Linking Text Sentiment to Public Opinion Time Series. In Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM 2010), 2010.
36. *Oren T., Dmitry D., Ari R.* ICWSM. A Great Catchy Name: Semi-Supervised Recognition of Sarcastic Sentences in Online Product Reviews. AAAI Conference on Artificial Intelligence, 2010.
37. *Pang B., & Lee L.* (2004). Sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of ACL 2004, pp. 271-278. New York: ACL Press.
38. *Pang B., Lee L.* Opinion Mining and Sentiment Analysis. Foundations and Trends in Information Retrieval, v.2 n.1-2, January, 2008, pp. 1-135.
39. *Pennebaker, J., Mehl, M., & Niederhoffer, K.* (2003). Psychological aspects of natural language use: Our words, our selves. Annual Review of Psychology, 54, pp. 547-577.
40. *Pollock, J. J., & Zamora, A.* (1984). Automatic spelling correction in scientific and scholarly text. Communications of the ACM, 27(4), pp. 358-368.
41. *Riloff, E., Patwardhan, S., & Wiebe, J.* (2006). Feature subsumption for opinion analysis. Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 440-448.

42. Sadikov Eldar, Aditya Parameswaran, and Petros Venetis. Blogs as predictors of movie success. In Proceedings of the Third International Conference on Weblogs and Social Media (ICWSM-2009), 2009.
43. *Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M.* (1966). The general inquirer: A computer approach to content analysis.Cambridge, MA: The MIT Press.
44. *Strapparava, C., & Valitutti, A.* (2004). Wordnet-affect: an affective extension of wordnet. In Proceedings of the 4th International Conference on Language Resources and Evaluation, pp. 1083-1086. Lisbon.
45. *Taboada, Maite, Brooke, Julian, Tofiloski, Milan, Voll, Kimberly, & Stede, Manfred.* (2011). Lexicon-based methods for sentiment analysis. Computational Linguistics, 37(2), pp. 267-307.
46. *Thelwall, M.* (2009). MySpace comments. Online Information Review, 33(1), pp. 58-76.
47. *Thelwall, M., & Buckley, K.* (2013). Topic-based sentiment analysis for the Social Web: The role of mood and issue-related words. Journal of the American Society for Information Science and Technology, 64(8), pp. 1608-1617.
48. *Thelwall, M., Buckley, K., & Paltoglou, G.* (2012). Sentiment strength detection for the social Web, Journal of the American Society for Information Science and Technology, 63(1), pp. 163-173.
49. *Thelwall, M., Buckley, K., Paltoglou, G. Cai, D., & Kappas, A.* (2010). Sentiment strength detection in short informal text. Journal of the American Society for Information Science and Technology, 61(12), pp. 2544-2558.
50. *Thelwall, M., Wilkinson, D., & Uppal, S.* (2010). Data mining emotion in social network communication: Gender differences in MySpace. Journal of the American Society for Information Science and Technology, 21(1), pp. 190-199.
51. *Tumasjan, Andranik, Timm O. Sprenger, Philipp G. Sandner, and Isabell M. Welpe.* Predicting elections with twitter: What 140 characters reveal about political sentiment. In proceedings of the International Conference on Weblogs and Social Media (ICWSM-2010), 2010.

52. *Turney, P. D.* (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th annual meeting of the Association for Computational Linguistics (ACL), July 6-12, 2002, Philadelphia, PA, pp. 417-424.
53. *Velikovich, L.; Blair-Goldensohn, S.; Hannan, K.; and McDonald, R.* 2010. The viability of web-derived polarity lexicons. In Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 777-785. ACL.
54. *Wiebe, J., Wilson, T., & Cardie, C.* (2005). Annotating expressions of opinions and emotions in language. Language Resources and Evaluation, 39(2-3), pp. 165-210.
55. *Wiebe, J., Wilson, T., Bruce, R., Bell, M., & Martin, M.* (2004). Learning subjective language. Computational Linguistics, 30(3), pp. 277-308.
56. *Wilson, T.* (2008). Fine-grained subjectivity and sentiment analysis: Recognizing the intensity, polarity, and attitudes of private states. University of Pittsburgh.
57. *Wilson, T., Wiebe, J., & Hoffman, P.* (2009). Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. Computational linguistics, 35(3), pp. 399-433.

Электронные ресурсы

58. Википедия. Анализ тональности текста: http://ru.wikipedia.org/wiki/анализ_тональности_текста
59. Программа SentiStrength: http://sentistrength.wlv.ac.uk/
60. Проект Lang-uk: http://lang.org.ua/uk/
61. Проект Pulse of the Nation: http://www.ccs.neu.edu/home/amislove/twittermood/
62. Утилита ReCal: http://dfreelon.org/utils/recalfront/
63. Хабрахабр. Автоматическое определение тональности текста (Sentiment Analysis): https://habrahabr.ru/post/263171/

64. Хабрахабр. Классификация данных методом опорных векторов: https://habrahabr.ru/post/105220/
65. Хабрахабр. Наивный Байесовский классификатор в 25 строк кода: https://habrahabr.ru/post/120194/
66. Хабрахабр. Обучаем компьютер чувствам (sentiment analysis по-русски): https://habrahabr.ru/post/149605/
67. Частотный словарь украинского языка: http://www.mova.info/freqcard2.aspx?l1=178&sl=pb4_all

Приложение 1. Исходные данные программы SentiStrength для украинского языка

EmotionLookupTable

азаза	2	безжурн*	2	боягуз*	-3
ахаха	2	беззакон*	-2	бояз*	-2
дрочи	-3	безнаді*	-3	боят*	-2
*срати	-2	безпереч*	1	бояч*	-2
*ссати	-2	безперіч*	1	боїм*	-2
уіі	3	безпечн*	2	боїт*	-2
хахаха	2	безпорадн*	-2	бридк*	-3
хуя	-4	безсерд*	-3	бридк*	-4
хіхіхі	2	безсил*	-2	брудн*	-2
іхіхі	2	безсором*	-3	буйн*	-2
3.14д*	-3	безстрашн*	2	буйств*	-2
3.14зд*	-3	безсумнівн*	1	бід	-2
аааа*	-3	безтурбот*	2	біда	-2
агрес*	-4	безхмарн*	2	бідам	-2
агресивн*	-3	безцінн*	2	бідами	-2
азарт*	3	безчесн*	-3	бідах	-2
акуратн*	2	бенкет*	3	біди	-2
алергі*	-2	бентеж*	1	бідола*	-2
ангельськ*	3	див	-3	бідою	-2
апетитн*	3	бидло	-2	біду	-2
афіген*	4	бизгрих*	2	біді	-2
ах*	3	била	-3	бій*	-2
ахах*	3	били	-3	біс	-2
б'*	-3	било	-3	біса	-2
багатолюд*	2	бити	-3	бісам	-2
бадьор*	2	благодат*	2	бісами	-2
бажан*	2	благодій*	2	бісах	-2
байст*	-3	благополуч*	2	бісе	-2
балда	-2	благоуспішн*	3	біси	-2
балдам*	-2	блажен*	3	бісит*	-3
балдах	-2	блеа*	-3	бісові	-2
балди	-2	бльо	-3	бісом	-2
балдою	-2	блюв*	-3	бісу	-2
балду	-2	бля*	-3	бісі	-2
балді	-2	бліа*	-3	бісів	-2
балді*	3	блін	-2	в'язниц*	-3
бариг*	-2	богин*	3	вб'*	-2
без журби	2	божествен*	4	вберег*	3
без суму	2	божеськ*	3	вбереж*	3
безграмотн*	-2	божечкі	3	вберіг*	3
безгріш*	2	боліт*	-3	вбив*	-2
бездоган*	4	борець	2	вбити*	-5
бездух*	-3	борот*	-2	вдал*	3
бездуш*	-3	борц*	2	вдари*	-2
безжал*	-4	бою*	-2	вдяч*	2

везуч* 2
величн* 3
веселі* 3
веселе* 3
весели* 3
весело 3
веселу* 3
вибач* 1
вибухов* -3
вигода 2
виграш* 3
вигідн* 2
виеб* -4
визначн* 3
вийоб * -4
викрав* -2
викрад* -2
викрас* -2
вина -2
вини -2
виною -2
вину -2
вину* -2
вині -2
виразн* 3
вирод* -3
вити -2
витончен* 3
вию* -2
вказу* -2
вмер* -2
вмира* -3
вогнян* -2
вонюч* -3
вонюч* -3
ворог* -3
ворож* -3
ворожк* 1
ворожіст* -2
воюв* -3
воюю* -3
впха* 2
враж* -2
враження 1
вредн* -2
врода 3
вродлив* 3
вріж* -3
врізав -3
врізал* -3
втом* -2
втоми* -2
втрач* -2
втіх* 3
вуха* 3
відбив -2
відбивати -2
відбил* -2
відбити -2
відваг* 3
відверн* -3
відлякув* -2
відмовля* 1
відмінн* 3
відпав* -2
відпад* -2
відрад* 3
відразл* -3
відразн* -3
відсмок* -3
відсмок* -3
відтягат* 3
відтягуват* 3
відчай* -4
відштовх* -2
відіб'* -2
вірніст* 2
вішат* -2
вії* 3
гандон* -4
ганеб* -2
ганьб* -3
гармоні* 3
гарна 3
гарне 3
гарни* 3
гарно* 3
гарну 3
гарні 3
гарній 3
гарячност* 3
гарячніст* 3
геніальн* 2
геро* 2
гидк* -3
гидлив* -2
гидливіст* -3
гидот* -3
глюч* -3
гнид* -3
гнил* -2
гнучк* 2
гнів* -3
гомофоб* -3
гоноровит* -3
горд* 3
гордовит* -2
грабіж* -2
грайлив* 3
граці* 3
гребан* -2
грозн* -3
груб* -3
груби* -2
грізн* -3
гріх* -4
гріш* -3
гімна -2
гімно -2
гірше -3
далбайоб* -3
даремн* -2
дарма -2
дбайлив* 2
дебіл* -3
денце -2
дибіл* -2
димляч* -3
днище -2
днюх* 2
днюш* 2
доблес* 3
добра 2
добраніч 2
добре 2
добрий 2
добрим 2
добрими 2
добрих 2
доброго 2
доброму 2
доброю 2
доброї 2
добру 2
добрі 2
добрій 2
добрім 2
довго -2
доведеться -2
довірят* 2
докуч* -2
долбойоб* -3
доплач* 2

допом*	3	загаджув*	-2	звіздабол*	-4
допомог*	1	загиб*	-5	згвалтув*	-3
досад*	-2	загину*	-5	згріш*	-3
доста*	-3	загроз*	-3	здох*	-2
достойн*	3	загрозлив*	-2	здохну*	-3
драту*	-3	загрузл*	-2	здраво	2
дратують	-2	загуб*	-3	зла	-3
дратує	-2	задерикуват*	2	зле	-3
дратівл*	-3	задзьобав*	-2	злий	-3
дриж*	-2	задовб*	-3	злим*	-3
дрож*	-2	задовбав*	-2	злих	-3
дроч*	-2	задоволенн*	3	злоб*	-3
дружн*	3	задовільн*	2	зловмисн*	-3
дубарь	-4	задра*	-3	зловісн*	-4
дур*	-3	задумлив*	-2	злого	-3
дуреп*	-3	задушевн*	3	злод*	-3
дурм*	1	зажур*	-2	злодія*	-4
дурш*	1	зазна*	-2	зломлен*	-4
духот*	-2	закохан*	2	злому	-3
дякс	2	закоху*	4	злочин*	-5
дяку*	3	заманлив*	2	злою	-3
діва	2	занепок*	-2	злої	-3
егоіс*	-3	запал*	3	злу	-3
ейфор*	4	запали*	-2	злі	-3
екстаз*	4	запальн*	2	злій	-3
елегантн*	3	запекл*	-4	злім	-3
ентузіа*	2	заплак*	-2	злісн*	-3
епічн*	4	заплач*	-2	зліст*	-3
жадан*	3	зарозуміл*	-2	зневаж*	-3
жалкува*	-2	зарозумілість*	-3	зневір*	-2
жаль	-2	засмут*	-2	знуща*	-4
жалюгідн*	-3	засмуч*	-2	зрад*	-3
жалі*	-3	заспок*	2	зраджув*	-2
жалібн*	-2	заспокій*	1	зрадник*	-3
жаліст*	-2	засра*	-2	зраду*	2
жарт*	3	затерт*	-2	зухвал*	-2
жах*	-3	затурбув*	-2	зухваліст*	-3
жахлив*	-4	затят*	-5	зірк*	2
жертв*	1	захват*	4	ибацц*	-2
життєрадіс*	3	захвилюв*	-2	йоб*	-3
життєрадісн*	2	захоплен*	3	йоб*	-4
жоп*	-3	захопленн*	4	йоп	-2
жорсток*	-4	зацікавл*	2	йопа*	-3
жостк*	-2	зашкод*	-2	йопт*	-3
журб*	-3	заєб*	-4	йопта	-2
з'їд*	-2	збентежи*	1	йопті	-2
забав*	2	збожеволі*	-2	йух	-3
забаганк*	-2	збс	3	кайф*	3
забан*	-2	збіговиськ*	-3	какаш*	-2
заборон*	-3	звеселянн*	3	капосн*	-3
завиваю*	-2	зворушлив*	3	капост*	-3

карколомн*	5	легкодух*	-3	мляя*	-2		
катастроф*	-2	ледач*	-3	мліст*	2		
катува*	-3	ледени*	1	могутн*	2		
кает*	-2	лестит*	2	молодец*	2		
каїт*	-2	лиховісн*	-4	молодц*	2		
квн	2	лицемір*	-2	морд*	-3		
кейф*	3	личать	3	моторошн*	-3		
кину*	-3	личимо	3	моторошн*	-3		
класн*	4	личите	3	мрійлив*	3		
кльов*	3	личити*	3	мрія*	2		
кмітлив*	2	личить	3	мудак*	-3		
коз*	-3	личиш	3	мудачін*	-2		
козел	-2	личу	3	мудень	-3		
козл*	-2	лол*	2	мудил*	-3		
козы*	1	лох	-2	мудо*	-3		
комплімент*	2	лох*	-3	мурмняв*	2		
комфорт*	2	лузер*	-2	мізерн*	-3		
конфуз*	1	люб'язн*	2	мімімішн*	3		
кончен*	-2	люб*	2	мімішн*	3		
коп	-2	любл*	3	мінорн*	-2		
копа	-2	лют*	-5	міф*	-2		
копу	-2	ляк*	-3	набрид*	-2		
корисн*	2	ляка*	-2	набридлив*	-2		
коха*	2	ляпас*	-3	наволоч*	-3		
кохан*	2	лід	1	нагруб*	-3		
кохаю*	4	ліки	2	надих*	2		
кошмар*	-3	лікув*	1	надихн*	3		
крадем*	-2	лінь	-2	надої*	-2		
крадемо	-2	майстерн*	3	наєб*	-2		
краса*	3	мандра*	1	найгірш*	-3		
красив*	3	маніяк*	-3	найздоровіш*	2		
красо*	3	марнот*	-3	найсмачніш*	4		
красти	-2	мат	-3	найсміливіш*	4		
кращ*	3	матам*	-3	найсприятливіш*	4		
краят*	-3	матах	-3	найсуворіш*	-3		
крається	-3	мати	-3	найталановитіш*	4		
криворук*	-2	матові	-3	найтепліш*	4		
крик*	-2	матом	-3	найцінніш*	4		
крич*	-2	мату	-3	найчарівніш*	4		
крута*	2	маті	-3	налагоди*	2		
крути*	2	матів	-3	насичув*	3		
кумедн*	3	мдаа*	-3	наскаржив*	-2		
курва*	-2	меланх*	-2	насолод*	2		
кіпіш*	-2	мерзенн*	-2	насолоджув*	3		
лага*	-3	мерзот*	-3	насуплен*	-2		
лайк*	2	мертв*	-3	натхн*	2		
лайн*	-2	мил*	2	натхненн*	3		
лал	2	мильн*	1	натіш*	2		
лапк*	2	млост*	2	нах	-2		
ласий	2	мля	-3	нахабн*	-2		
ласк*	3	млять	-3	нахер	-2		

нахуй	-4		нитт*	-2		охіре*	-4
неабияк*	2		ниціст*	-3		очешує*	3
небезпеч*	-3		норм*	2		очман*	3
невдач*	-2		нуди*	-3		ощаслив*	2
невдячн*	-2		нудист*	1		панік*	-3
невиразн*	-2		нудн*	-2		паплюжи*	-4
невмотивован*	-3		нудот*	-3		паскуд*	-3
негарн*	-2		нудь*	-3		пацюк*	-3
недобр*	-3		нєфіг*	-3		педик*	-2
недоброзичлив*	-2		ніжа*	2		педіг*	-2
недовірлив*	1		ніжи*	2		пекло	-3
недовірлив*	-3		ніжн*	2		перебит*	-3
недруж*	1		ніким	-2		перегар*	-2
неебу	-3		нікчемн*	-3		переляк*	-2
незабутн*	2		нірван*	3		перемог*	2
незадоволен*	-2		нііпет	-3		переможж*	2
незамінн*	2		облам*	-3		печал*	-2
нездійснен*	-2		обман*	-2		пздец*	-4
незручн*	-3		обож*	3		пздц	-3
незрівнянн*	3		обрадува*	3		пздц*	-4
неймовірн*	3		ображат*	-3		пидор*	-3
нелюдим*	-2		образ	1		пизд*	-4
немилосердн*	-4		образ*	-2		пиздобол*	-3
неможлив*	-2		образи*	-4		пипец	-3
ненави*	-4		образлив*	-4		писк*	-3
необходітельн*	-2		обурен*	-2		писок	-3
неординарн*	2		обурю*	-3		пистец	-3
неповноцінн*	-2		обурюва*	-2		пихат*	-2
непоган*	2		обіймм*	2		пихатіст*	-3
неправ*	-2		огидн*	-3		пиша*	2
непривітн*	-2		огидн*	-4		пишна	3
непристойн*	-3		одружи*	3		пишний	3
неприхильн*	1		одужа*	2		пишно	3
неприязн*	-3		оживл*	2		плакав*	-2
неприємн*	-2		озлоби*	-4		плакав*	-3
нерасположен*	1		озлоблен*	-3		плакал*	-2
нервов*	-2		окупант*	1		плакат*	-3
нерозумн*	-2		омріян*	3		плюват*	-3
несамовит*	-3		оптимізм*	2		плюс	2
несмілив*	1		оптимістичн*	2		плющит*	-3
неспок*	-3		оскаженіл*	-3		плюють	-3
неспокійн*	-2		остерв*	-4		плює	-3
несправедлив*	-2		от'єб*	-5		плюєш	-3
нетовариськ*	-2		отетер*	3		пля	-2
неуравновешен*	-2		офіген*	5		пляя*	-3
нехтува*	-2		офігів*	2		пнх	-4
нещадн*	-4		офіційн*	2		поб'*	-3
нещасн*	-3		охрен*	-2		побалд*	3
нещасн*	-3		охуде*	-2		поби*	-3
неясн*	-2		охуе*	-4		повага	2
низьк*	-3		охуенн*	3		повесел*	4

повія -3
поган*-2
поганя* 1
погож* 2
погрози* 1
погріш* -4
подавл* -3
подар* 1
подоба 1
подоба* 2
подяк* 2
пожалі* -2
пожвавл* 2
позитив* 2
покаран* -2
покарат* -2
покаят* -2
покуй* -4
покій* 1
полохл* -3
полум* -3
полюб* 3
помер* 1
помилк* -2
помилк* -2
поносн* -4
понудьг* -2
понур* -3
попсятін* -2
пороч* -3
порочн* -2
поруш* -2
посварити* -2
поскаржив* -2
посмішк* 2
посрат* -4
постражд* -2
посіпак* -3
потворн* -2
потрібн* 1
потіх* 2
потішн* 2
пох -2
похвальн* 2
похер* -2
похмур* -3
похуй* -4
похую* -4
поціл* 3
пошкодува* -2
ппц -3
правильн* 1
представ* 1
прекрасн* 3
престижн* 3
приваблив* 3
пригнічен* -4
приголомшлив* 4
прикрист* -2
прикро -2
прикріст* -3
приниж* -4
приниз* -4
пристойн* 2
прихильн* 2
приємн* 3
проб'* -2
пробач* 1
проби* -2
провал* -3
провин* -3
прогав* -2
проеб* -2
проебав* -3
проклина* -4
проклят* -4
проступ* -2
проти 1
противн* -2
процвіт* 3
псіх* -2
підар* -3
підераст* -3
підйоб* -2
підлаб* -2
підли* -2
підліст* -3
піднесені* 3
підор* -3
підпал* -4
підр -3
підступн* -2
підступн* -3
підходящ* 2
пізд* -4
пісд* -4
рад 2
рада 2
радий 2
радост* 3
радіст* 3
радіст* 3
разюч* 3
раюванн* 2
ревіти -2
ремствуван* -2
респект* 3
ржач* 2
розбещен* -2
розбит* -3
розбурхан* -2
розваг* 3
роздосад* -3
роздратува* -3
розкаюва* 1
розкіш* 3
розлад* -3
розлют* -5
розлюти* -3
розлюч* -5
розпещен* -2
розпіздя* -2
розсердж* -3
розумн* 2
розчарова* -3
розчуленн* 3
романтич* 2
руйнац* -3
садист* -3
самогуб* -5
самолюб* -2
самоповаг* 2
сволот* -3
свят* 3
сексуальн* 3
сердит* -3
симпатич* 2
симпаті* 2
ска -2
скажен* -4
сказ -4
скази -4
сказові -4
сказу -4
сказу -4
сказів -4
скандаль* -2
скарг* -2
скиглити -2
скопитився -2
скорб*-3
скорботн* -4
скот -3

скота -3
скоте -3
скотові -3
скотом -3
скоту -3
скоті -3
скромн* 1
скрушен* -3
скуч* -2
скучер* 1
слаб* -2
славн* 2
сльоз* -2
смакува* 2
смачн* 2
смерт* -4
смокт* -2
смокч* -2
смут* -2
сміттє* -2
сміх 2
смішн* 3
сонечк* 4
сонячн* 3
сором'язлив* 1
сором* -4
соромит* 1
соромн* -3
спасибі 2
спека 3
сподоб* 2
спок* 3
спокій* 1
справедлив* 3
сприянн* 3
сприят* 3
сприятлив* 4
співчув* -3
співчутт* -3
срак* -5
сран* -3
стерв* -3
стопудов* 3
стражд* -3
страх* -4
страхітт* -3
страш* -3
стрьомн* -2
ступор* -2
сувор* -3
сук -3
сука -3
сукам* -3
суках -3
суки -3
суко -3
сукою -3
суку -3
сум -2
суми -2
сумові -2
сумом -2
суму -2
суму* -2
сумі -2
супер* 3
суук* -3
сууук* -3
суці -3
сучар* -4
сучий син -3
сучк* -3
сучков* -2
суїцид* -5
схвален* 2
схвальн* 2
схвалюв* 2
сюрприз* 2
тааа* 2
тактактак* 3
тактовн* 2
талант* 2
тварина -3
тварь -2
тварюк* -4
тероріз* -4
тлін* -3
томлінн* -2
тоскн* -2
траур* -5
тремті* -2
трепет* -2
трепотінн* -2
треш* -2
триво* -3
тріумфув* 5
туга -2
туго* -2
туж* -3
тужн* -2
тужур* 1
тупит* -2
тупиш -2
тупособразітєльн* -3
турбува* -2
тьмян* -2
тішит* 5
тішу* 5
убог* -3
удач* 2
улюблен* 3
унікальн* 3
упокій* 1
ура 3
урочист* 3
уславлен* 4
успіх* 3
успішн* 3
утіх* 3
ух 2
ущербн* -4
уєба* -3
фак -3
фейл* -3
фестивал* 3
флірт* 2
фотогенічн* 2
фурор* 4
ффу* -2
фіг -3
фіга -3
фігн* -2
фієст* 2
халяв* 3
хворі* -3
хд* 2
хд* 2
хер -2
хера -2
херу -2
хизуват* 1
хитр* -3
хня -2
хоробр* 2
хорош* 3
хохо 2
хочу 2
хрін* -2
хуй* -4
хуйзна-як -2
хуль -2
хулі -2
хууй -4

хуєв*	-3	чудесн*	3	янгол*	3		
хуєсос*	-4	чудов*	3	яскрава	3		
целюл*	-4	чудовиськ*	-4	яскравий	3		
целюло*	1	чудовищ*	-4	яскраво	3		
цікав*	2	шалав*	-2	яскраві	3		
цілитель	2	шалені*	-2	ясна	2		
цілувати	3	шаленіст*	-3	ясний	2		
цінн*	2	шана	3	ясно	2		
цінник*	1	шану*	3	єб*	-4		
чарівн*	2	шедевр*	4	єбать	-3		
чван*	-3	шикарн*	3	єбашив*	-3		
чванит*	1	шкода	-2	єбе	-3		
чванлив*	1	шлендр*	-4	єбт	-3		
чванливи*	-2	шльондр*	-3	єбуч*	-3		
чеее*	-2	шляхетн*	3	ідеальн*	2		
чесн*	2	шмара	-4	ідіот*	-3		
чинуш*	-3	шняг*	-2	імбеціл*	-3		
чистота	2	шулер*	-2	імбіцил*	-3		
чмо	-2	щаслив*	3	іннах	-2		
чмошнік*	-2	щаст*	4	іпацц*	-2		
чмошніц*	-2	щедро	3	іпохондр*	-2		
чорт*	-3	щиро	3	іпёт	-3		
чотк*	2	якісн*	2	їб*	-3		

BoosterWordList

абсолютно	2	зобов'язаний	-1	робив	1		
би	1	зробив	1	сама	1		
був	-1	миленько	3	самий	1		
буд	-1	може	1	самого	1		
важко	-2	можна	-1	сліпуче	3		
варто	-1	можу	1	сліпучо	3		
вже	1	міг	1	справді	1		
виключно	2	на жаль	-2	справжній	1		
виразно	1	надзвичайно	3	супер	2		
вкрай	3	наскрізь	2	так	1		
воістину	1	настільки	1	трохи	-1		
гребанние	-3	невимовно	3	трошки	-1		
дідьків	-2	неймовірно	2	убивчо	2		
дивно	3	нереально	2	удачі	2		
досконало	2	несказанно	3	ультра	2		
дуже	2	особливо	1	феноменально	3		
дійсно	1	певно	1	ціле	1		
ефектно	3	повинен	-1	чортів	-2		
жахливо	2	повинн	-1	чудово	3		
загалом	1	повністю	1	ще	-2		
занадто	-2	погано	-2	щиро	2		
злегка	1	послідовно	1	як	1		
значить	1	приголомшливо	3				
зобов'язан	-1	просто	1				

NegatingWordList

без
не
нема
немає
ні
ніколи

QuestionWords

де
коли
навіщо
хто
чому
що
як

Приложение 2. Примеры оцененных программой SentiStrength твитов

оценка экспертов		текст	оценка программы		
1	2	таке враження, що технічний прогрес обминув наш благодатний край? https://t.co/rLcO5PmOwE	2	-1	таке[0] враження[0] що[0] технічний[0] прогрес[0] обминув[0] наш[0] благодатний[1] край[0] [[Sentence=-1,2=word max, 1-5]] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/rLcO5PmOwE[0] [[Sentence=-1,1=word max, 1-5]][[[2,-1 max of sentences]]]
1	1	Якщо розкласти перемогу на найменші деталі, її не відрізнити від поразки	2	-1	Якщо[0] розкласти[0] перемогу[1] на[0] найменші[0] деталі[0] її[0] не[0] відрізнити[0] від[0] поразки[0] [[Sentence=-1,2=word max, 1-5]][[[2,-1 max of sentences]]]
1	1	Світовий юніорський рекорд у Запоріжжі від Аліни Шух! https://t.co/pxJlroigqR	1	-1	Світовий[0] юніорський[0] рекорд[0] у[0] Запоріжжі[0] від[0] Аліни[0] Шух[0] [[Sentence=-1,1=word max, 1-5]] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/pxJlroigqR[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
1	2	Ще швагро розповідав, як він сам вирішив розбагатіти. Вирив три котловани, мало не получив п..зди від місцевих і не знайшов жодного камінця.	1	-1	Ще[0] швагро[0] розповідав[0] як[0] він[0] сам[0] вирішив[0] розбагатіти[0] [[Sentence=-1,1=word max, 1-5]] Вирив[0] три[0] котловани[0] мало[0] не[0] получив[0] п[0] [[Sentence=-1,1=word max, 1-5]] зди[0] від[0] місцевих[0] і[0] не[0] знайшов[0] жодного[0] камінця[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
1	3	Включився мозг. Стояночка. Об' ясність мені, шо вчора проісходило і якого чорта це було вообще?	1	-3	Включився[0] мозг[0] [[Sentence=-1,1=word max, 1-5]] Стояночка[0] [[Sentence=-1,1=word max, 1-5]] Об'[0] ясність[0] мені[0] шо[0] вчора[0] проісходило[0] і[0] якого[0] чорта[-2] це[0] було[0] вообще[0] [[Sentence=-3,1=word max, 1-5]][[[1,-3 max of sentences]]]
1	1	Трамп сказав у п'ятницю, що він тільки на ранніх стадіях розгляду скасування санкцій щодо раші. https://t.co/76okXGCFij via @Reuters	1	-1	Трамп[0] сказав[0] у[0] п'ятницю[0] що[0] він[0] тільки[0] на[0] ранніх[0] стадіях[0] розгляду[0] скасування[0] санкцій[0] щодо[0] раші[0] [[Sentence=-1,1=word max, 1-5]] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/76okXGCFij[0] via[0] @Reuters[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
1	3	9 кіл пекла - це не в Данте, це в Укрзалізниці.	1	-1	9[0] кіл[0] пекла[0] це[0] не[0] в[0] Данте[0] це[0] в[0] Укрзалізниці[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
1	1	Адвокат Януковича хоче, щоб він сам пересвідчився у власній зраді https://t.co/ntymFLRrZw	1	-3	Адвокат[0] Януковича[0] хоче[0] щоб[0] він[0] сам[0] пересвідчився[0] у[0] власній[0] зраді[-2] https[0] ://t[0] [[Sentence=-3,1=word max, 1-5]] co/ntymFLRrZw[0] [[Sentence=-1,1=word max, 1-5]][[[1,-3 max of sentences]]]
1	3	Недавно люди отримали свободу і нам дозволили відкрити рот. Як дивно, нам дали те, що завжди	1	-1	Недавно[0] люди[0] отримали[0] свободу[0] і[0] нам[0] дозволили[0] відкрити[0] рот[0] [[Sentence=-1,1=word max, 1-5]] Як[0] дивно[0] нам[0] дали[0] те[0] що[0] завжди[0] належало[0] нам[0] [[Sentence=-1,1=word max, 1-5]] Коли[0] вони[0]

		належало нам. Коли вони дозволять нам дихати?			дозволять[0] нам[0] дихати[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
1	3	мені вже надоїло то слухати від всіх	1	-3	мені[0] вже[0] надоїло[-1][-1 LastWordBoosterStrength] то[0] слухати[0] від[0] всіх[0] [[Sentence=-3,1=word max, 1-5]][[[1,-3 max of sentences]]]
2	1	Уууу ууууу саме ту зустріну іншу що вірна /есть исключения украинских песен которые я слушаю/	2	-1	Уууу/Ууу[0] ууууу/уу[0][+0.6 MultipleLetters] саме[0] ту[0] зустріну[0] іншу[0] що[0] вірна[0] /есть[0] исключения[0] украинских[0] песен[0] которые[0] я[0] слушаю/[0] [[Sentence=-1,2=word max, 1-5]][[[2,-1 max of sentences]]]
2	2	Цей біолог просто космос.Кожен ранок дивлюсь його в "Сніданок з 1 1" і тупо ору.Якийсь Гліб.Цікаво,це він в житті такий намаханий,чи гонить?	2	-1	Цей[0] біолог[0] просто[0] космос[0] [[Sentence=-1,1=word max, 1-5]] Кожен[0] ранок[0] дивлюсь[0] його[0] в[0] Сніданок[0] з[0] 1[0] 1[0] і[0] тупо[0] ору[0] [[Sentence=-1,1=word max, 1-5]] Якийсь[0] Гліб[0] [[Sentence=-1,1=word max, 1-5]] Цікаво[1] це[0] він[0] в[0] житті[0] такий[0] намаханий[0] чи[0] гонить[0] [[Sentence=-1,2=word max, 1-5]][[[2,-1 max of sentences]]]
1	1	Фонд гарантування вкладів фізосіб на підставі рішення Нацбанку від 26 січня 2017 року про віднесення ПАТ "Фортуна-… https://t.co/d0Cc5LhKKr	1	-1	Фонд[0] гарантування[0] вкладів[0] фізосіб[0] на[0] підставі[0] рішення[0] Нацбанку[0] від[0] 26[0] січня[0] 2017[0] року[0] про[0] віднесення[0] ПАТ[0] Фортуна[0] -…[0] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/d0Cc5LhKKr/co/d0Cc5LhKr[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
5	1	а Кріс красавчік, як він подмігую бооожечкі якій афігенний	4	-1	а[0] Кріс[0] красавчік[2] як[0] він[0] подмігую[0] бооожечкі/боожечкі[0] якій[0] афігенний[3] [[Sentence=-1,4=word max, 1-5]][[[4,-1 max of sentences]]]
1	1	Він Пам'ятник Собі Побудував Нерукотворний! - https://t.co/mvtSNpA9Ed	1	-1	Він[0] Пам'ятник[0] Собі[0] Побудував[0] Нерукотворний[0] [[Sentence=-1,1=word max, 1-5]] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/mvtSNpA9Ed[0] [[Sentence=-1,1=word max, 1-5]][[[1,-1 max of sentences]]]
1	2	Зараз він заплаче! "А ви говорите – газ дорогий". Що пишуть соцмережі про зарплату Коболєва https://t.co/cTljFvGTU0 #Нафтогаз	1	-2	Зараз[0] він[0] заплаче[-1] [[Sentence=-2,1=word max, 1-5]] А[0] ви[0] говорите[0] –[0] газ[0] дорогий[0] [[Sentence=-1,1=word max, 1-5]] Що[0] пишуть[0] соцмережі[0] про[0] зарплату[0] Коболєва[0] https[0] ://t[0] [[Sentence=-1,1=word max, 1-5]] co/cTljFvGTU0[0] #Нафтогаз[0] [[Sentence=-1,1=word max, 1-5]][[[1,-2 max of sentences]]]