

# Метод определения эмоций в текстах на русском языке

A method of sentiment analysis  
in Russian texts

**Пазельская А. Г.** (pazelskaya@i-teco.ru)

**Соловьев А. Н.** (a.solovyev@i-teco.ru)

ЗАО «Ай-Текс», Москва, Россия

В работе рассматриваются методы автоматического определения эмоциональной составляющей (тональности) в тексте и описывается опыт осуществляемой в данный момент практической реализации системы для текстов СМИ на русском языке, в основе которой лежат словари лексической тональности и набор комбинаторных правил объединения отдельных слов и словосочетаний.

В работе впервые предложен метод определения тональности, основанный на предикационных отношениях в пропозиции. В связи с этим нами предложена классификация глаголов в зависимости от их эмоционального воздействия и местоположения объекта тональности.

## 1. Эмотивная составляющая в тексте

В данной статье рассматривается один из методов определения эмоционального компонента в тексте. Эта задача относится к обширному кругу задач анализа и обработки различных функций коммуникации на естественных языках. Сегодня в современных информационных технологиях широко применяются системы обработки коммуникационной (или информационной) и метаязыковой функций<sup>1</sup> коммуникации. Наряду с этим возникает необходимость обработки и других функций: фатической, аппеллятивной и эмотивной (в том числе оценочной). Информационная функция коммуникации применяется при взаимодействии человека с компьютером, когда нужно получить или уточнить необходимую информацию (например, справочные системы). В автоматических системах перевода используется метаязыковая функция — кодирование и изоморфное преобразование языковой информации. На фатической функции основываются различные развлекательные системы, поддерживающие диалог с пользователем, в том числе с применением речевых технологий (см., например, [Соловьев и др., 2003]). Эмотивную функцию коммуникации

---

<sup>1</sup> Согласно классификации функций коммуникации, данной Р. Якобсоном в своей работе «Лингвистика и поэтика» [Якобсон, 1975].

пытаются использовать в автоматических системах оценки и сравнения объектов, например, новых продуктов и брендов известных компаний, для выявления отношения людей к событиям в политической жизни страны и т. п.

Эмоциональная составляющая коммуникации пока не столь активно применяется в системах обработки текстовой информации не только ввиду трудностей выделения «нужной» (т. е. относящейся к рассматриваемому объекту) эмоциональной лексики в текстах, но и сложности определения самого эмотивного пространства, количества и состава его измерений<sup>2</sup>. К сожалению, теория эмоций в лингвистике еще недостаточно развита.

Исследования в области теории лингвистических эмоций начались не так давно. В 50-х годах прошлого века Чарльз Осгуд с помощью метода семантического дифференциала пытался определять эмотивное пространство различными наборами парных слов [Осгуд и др., 2007]. В настоящее время исследования эмоций лежат в основном в области психологии, нейрофизиологии и психолингвистики. В лингвистике разработаны психометрические инструменты и методы для таких исследований (см., например, OpinionFinder [Wilson & al. 2005] или Profile of Mood States (POMS-bi) [Norcross & al., 2006]).

В современных системах автоматического определения эмоциональной оценки текста чаще всего используется одномерное эмотивное пространство: позитив-негатив, то есть хорошо-плохо. Однако известны успешные случаи использования и многомерных пространств [Bollen & al., 2010]. Более подобный обзор современного состояния в области анализа тональности текста представлен в книге [Pang & Lee, 2008].

В нашем методе при эмоциональной оценке рассматриваемого текста мы используем эмотивное пространство, содержащее негативную-позитивную составляющую плюс силу эмотивности.

## 2. Понятие лексической тональности и тональности предложения

Эмоциональная оценка, выраженная в тексте, также называется тональностью, или сентиментом текста (от англ. sentiment — чувство; мнение, настроение). Эмоциональная составляющая, выраженная на уровне лексемы или коммуникативного фрагмента<sup>3</sup>, называется лексической тональностью (или лексическим сентиментом). Тональность текста в целом определяется лексической тональностью составляющих его единиц и правилами их сочетания.

---

<sup>2</sup> Например, при оценке электронных продуктов *большое разрешение монитора* — это хорошо, но *большой вес* — плохо; прилагательное *большой* ведет себя по разному с точки зрения эмоциональной оценки, в зависимости от того, о чем идет речь.

<sup>3</sup> «Коммуникативные фрагменты (КФ) — это отрезки речи различной длины, которые хранятся в памяти говорящего в качестве стационарных частиц его языкового опыта и которыми он оперирует при создании и интерпретации высказываний». [Гаспаров, 1996]. Коммуникативный фрагмент обычно больше, чем слово, но меньше, чем предложение.

Автоматическое определение тональности текста подразумевает выделение тех фрагментов текста, которые выражают позитивную или негативную эмоциональность по отношению к объекту эмоциональной оценки (объекту тональности). Таким объектом может быть имя собственное, название продукта, организации, услуги, профессии и т. п., по отношению к которому анализируется текст. Объект эмоциональной оценки может быть задан как один в целом для текста (с учетом его синонимических и анафорических употреблений), так и определяться в предложениях как любое имя собственное или даже нарицательное.

Таким образом, тональность текста определяется тремя факторами: 1) субъект тональности; 2) собственно тональная оценка (позитив/нейтрально/негатив); 3) объект тональности. Под субъектом тональности подразумевается автор статьи (автор цитаты, прямой или косвенной речи), под объектом тональности — тот, о ком он высказывается и под тональной оценкой — эмоциональное отношение автора к такому объекту.

### 3. Методы определения тональности текста

Существуют три основных метода определения тональности текста.

1. Анализ текста методами векторного анализа (часто с применением n-граммных моделей), сравнение с ранее размеченным эталонным корпусом по выбранной мере близости и отнесение (классификация) текста к негативу или позитиву на основании полученного результата сравнения.
2. Поиск эмотивной лексики (лексической тональности) в тексте по заранее составленным тональным словарям (спискам паттернов) с применением лингвистического анализа. По совокупности найденной эмотивной лексики текст может быть оценен по шкале, отражающей количество негативной и позитивной лексики. Этот метод может использовать как списки паттернов, подставляемые в регулярные выражения, так и правила соединения тональной лексики внутри предложения.
3. Смешанный метод (комбинация первого и второго подходов).

Первый метод (см., например, [Pang & al., 2002; Pang & al., 2005; Gamon, 2004]) работает достаточно быстро, но требует наличия предварительно размеченного эталонного корпуса, на основе которого происходит обучение алгоритма сравнения. Существенными недостатками такого подхода оказываются увеличение трудоемкости и ограничение разнородности корпуса (т. е. неполнота лексического покрытия), что приводит к потере точности. К тому же данный метод не позволяет провести глубокий анализ текста, то есть выявить и показать эмотивность на уровне предложения.

Второй метод [Nasukawa, 2003; Yi, 2003] не менее трудоемок в составлении тональных словарей (или получения списка тональных паттернов),

но в сочетании с синтаксическим и морфологическим анализом более гибок: он позволяет не только показать цепочки тональной лексики, но и получить синтаксически корректные эмоциональные выражения. При хорошем наполнении тональных словарных списков этот метод позволяет достичь хорошей полноты (покрытия эмотивной лексики).

Недостаток этого метода в том, что с помощью него сложно дать количественную оценку негативности-позитивности текста. Чтобы избежать недостатков первого и второго метода, используют смешанный подход [Prabowo & al., 2009; König, 2006], частично включающий в себя два первых.

Мы опишем методику и опыт использования второго метода определения тональности текста с использованием правил объединения слов в цепочки и определения тональности у объекта на основе предикационных отношений в пропозиции. Создаваемая нами система предназначена для обработки новостных текстов общероссийских СМИ.

#### **4. Определение тональности с использованием тональных словарей и лингвистического анализа**

Анализ тональности текста, реализуемый нами в настоящий момент, состоит из нескольких этапов. Сначала обрабатывает отдельный лингвистический модуль, автоматически производящий морфологический анализ текста, лемматизацию всей лексики и определяющий части речи каждого слова, его морфологические характеристики (падеж, лицо, число, активность-пассивность для глаголов), роль этого слова в предложении (для существительных: подлежащие, обстоятельство, дополнение; для глаголов: причастие, деепричастие, глагол; и др.), его тип (например, для существительных: физическое лицо, юридическое лицо, географическое название и др.).

Затем все слова (существительные, глаголы, прилагательные и наречия) и некоторые словосочетания (коллокации) размечаются по заранее подготовленным словарным спискам тональной лексики. Каждому слову приписывается два атрибута, указывающие на тональность и/или силу тональности. Если слово не нашлось в списках тональной лексики, то оно считается нейтральным.

После этого запускается первичный синтаксический анализ: слова и словосочетания объединяются в тональные цепочки, в предложении выделяются субъект, предикат и объект, идентифицируются причастные и деепричастные обороты, подчинительные предложения, анафорические связи и пр. Естественно, не каждое предложение русского языка можно представить в виде триады субъект-предикат-объект. Учитываются также безличные, неопределенно-личные и обобщенно-личные предложения, предложения с нулевой формой глагола, сказуемые, выраженные неглагольной формой.

На последнем этапе в предложении выделяется объект тональности и определяется его сентимент в зависимости от местоположения и роли этого объекта в предложении.

## 4.1. Тональные словари

Таким образом, необходимое условие для анализа тональности — составление словарного списка тональной лексики. Мы использовали тональные словари, разделенные по четырем частям речи (существительные, глаголы, прилагательные и наречия), плюс глагольные и неглагольные коллокации<sup>4</sup>. Использование коллокаций было вызвано тем, что далеко не все сочетания слов при объединении их по общим правилам дают в результате правильный сантимерт (например, *общество с ограниченной ответственностью, взрыв смеха* и пр.).

Все части речи разделяются на разные подклассы в зависимости от лексической тональности. Например, словарь глаголов состоит из одиннадцати подклассов (см. подробнее соответствующий раздел). Тональные словари заполняются экспертно; в начале работ был размечен лексический сантимерт наиболее частотных слов разных частей речи, извлеченных из составленного специально для этой цели на основе информационных русскоязычных порталов Интернета корпуса текстов СМИ (около 100 млн. словоупотреблений). В процессе тестирования и отладки системы тональные списки постоянно пополняются и сейчас содержат более 15 000 тональных слов и коллокаций. В словари попадают только слова и словосочетания, несущие какую-либо тональность или усиливающие тональность связанных с ними единиц.

Каждое слово или словосочетание может попасть только в один из классов по частям речи и тональности. Естественно, при таком подходе мы сталкиваемся с проблемой омонимии (одно и то же слово может иметь разный сантимерт или даже принадлежать разным частям речи). Эту проблему мы частично снимаем с помощью увеличения списка коллокаций и учета глагольного управления (ср. *болеть за что-л.* и *болеть чем.-л.*). Слова, тональность которых зависит от тематики текста, размечались согласно тому, в каком качестве их эмотивность или сила эмотивности чаще употребляется в корпусе СМИ.

### 4.1.1. Неглагольные лексемы и коллокации

Наречия, прилагательные и неглагольные коллокации делятся на позитивные, негативные и усиливающие эмоциональность, то есть такие слова или словосочетания, которые сами по себе не несут сантимерта, но при этом могут усиливать эмоциональность того, к чему присоединяются (например, наречия *круто, ужас*; прилагательные *эксклюзивный, потрясающий* и коллокации *коренным образом, решающая роль*). Сила тональности определялась экспертами по трехбалльной шкале.

Имена существительные также могут быть позитивными (например, *благотворительность* или *зарплата*) и негативными (*налог* или *война*). Однако не все существительные имеют однозначную эмоциональную нагрузку, тональность многих зависит от окружения. Поэтому целесообразно вводить классы потенциально негативных и потенциально позитивных слов — так,

---

<sup>4</sup> Под коллокациями понимались любые устойчивые и достаточно часто встречающиеся сочетания слов, как с идиоматическим значением, так и с неидиоматическим.

потенциально позитивные слова позитивны в позитивном окружении и нейтральны во всех остальных. Например, слово *план* само по себе не несёт в себе тональности, но сочетание *план по выходу из кризиса* должно давать позитив.

Особую роль играют отглагольные существительные: они могут менять тональность следующего за ним существительного. Например, отглагольное существительное *прекращение* меняет её на противоположную. Если за ним следует позитивная цепочка связанных существительных, например, *прекращение поставок угля*, то объединенная цепочка будет негативной. Если за данным отглагольным существительным следует негативная цепочка, например, *прекращение военных действий*, то в целом новая цепочка получит позитив. Поэтому отглагольные существительные выделялись в два отдельных класса: меняющие тональность зависящего от них слова (как *прекращение* или *спад*) и сохраняющие её (*рост* или *проведение*).

#### 4.1.2. Глагольные лексемы и коллокации

Особое внимание при разработке мы уделили классификации и тональной разметке глаголов. В нашем методе именно предикация (элементарная единица текста, состоящая из глагола и его зависимых) является ключевой составляющей при определении тональности объекта. Иными словами, тональность определяется (в общем случае) тремя составляющими: тональностью самого объекта, действием, производимым объектом или над объектом, и тональностью остальных участников описываемой ситуации.

Любое упоминание объекта в предложении характеризуется двумя параметрами: его окружением и его ролью относительно глагола. Каждый из этих параметров может как влиять, так и не влиять на итоговую тональность объекта в предложении — это определяется глаголом. Соответственно, в зависимости от влияния этих двух параметров на тональность объекта и тональности глагола как такового мы выделили восемь классов глаголов:

- 1 и 2 класс — негативные и позитивные глаголы, определяющие тональность объекта в зависимости от окружения, но независимо от его роли (негативные *уносить, стереть, освободить от*; позитивные *защищать, предрекать, болеть за*);
- 3 и 4 класс — негативные и позитивные глаголы, определяющие тональность объекта независимо от окружения, но в зависимости от его роли (например, глаголы *сдаться* и *проиграть* приписывают негатив субъекту и позитив объекту, а глаголы *обуздать* и *повергнуть*, наоборот, приписывают позитив субъекту и негатив объекту);
- 5 и 6 класс — негативные и позитивные глаголы, определяющие тональность объекта в зависимости от его окружения и роли (в основном в эти классы вошли возвратные глаголы; примеры негативных глаголов: *жаловаться, испугаться, замерзать*; позитивных глаголов: *окупаться, влечь, согреться*);
- 7 и 8 класс — чисто негативные и чисто позитивные глаголы, определяющие тональность объекта вне зависимости от его роли и окружения (например, *расследовать* и *улучшать* всегда приписывают позитив, а *грабить* и *злоупотреблять* — негатив);

Кроме того, отдельно выделяются три дополнительных класса глаголов:

- 9 класс — глаголы, соединяющие или приравнивающие тональность объекта и субъекта (так называемые связочные глаголы: *являться, олицетворять, относиться*).
- 10 и 11 классы — позитивные и негативные глагольные коллокации, например, негативные *наложить руки, освободить от должности, пробиловать насквозь* и позитивные *поразить противника, заострить внимание, заливаться смехом*.

Списки глаголов составлялись с учетом глагольного управления: глаголы, тональность которых менялась в зависимости от глагольного управления, попадали в разные классы (например, *высказаться за* и *высказаться против*). Также каждому глаголу и глагольной коллокации по трехбалльной шкале была приписана сила тональности (это наиболее важно для 7–8-го классов и списков коллокаций).

## 4.2. Правила сочетаемости

Следующий подготовительный этап тонального анализа текста — составление правил сочетаемости лексем и коллокаций. Слова и словосочетания соединяются по этим правилам между собой, причём сначала объединяются соседние неглагольные элементы, затем они присоединяются к глаголу, определяя, таким образом, сантимерт внутри предикации (простого предложения). Не все правила можно задать корректно: например, сочетание негативного и позитивного существительного в общем случае не определено. Наиболее частотные из таких словосочетаний включались в списки коллокаций, остальные обрабатывались в зависимости от глагола, возглавляющего предложение.

Правила представляют собой комбинации различных членов предложения между собой. Учитываются слова-инверторы, например, *не, нет, без, вне* и пр.

Разрешаются анафорические ссылки, выраженные местоимениями и местоименными словами. Сложные предложения разбиваются на простые, некоторые типы придаточных предложений включаются в родительское, причастные обороты присоединяются к определяемому слову, деепричастные — к субъекту родительского предложения. Придаточные предложения определительного типа с разрешенной анафорией соотносятся с определяемым словом.

В итоге предложение приводится к одному из типов синтаксической структуры из субъекта, предиката и объекта, где каждый член структуры в общем случае представлен цепочкой словоформ с определенной тональностью.

В случае нулевого глагола тональность определялась по тональному окружению объекта, его качественным признакам.

## 4.3. Определение сантимерта по отношению к объекту тональности

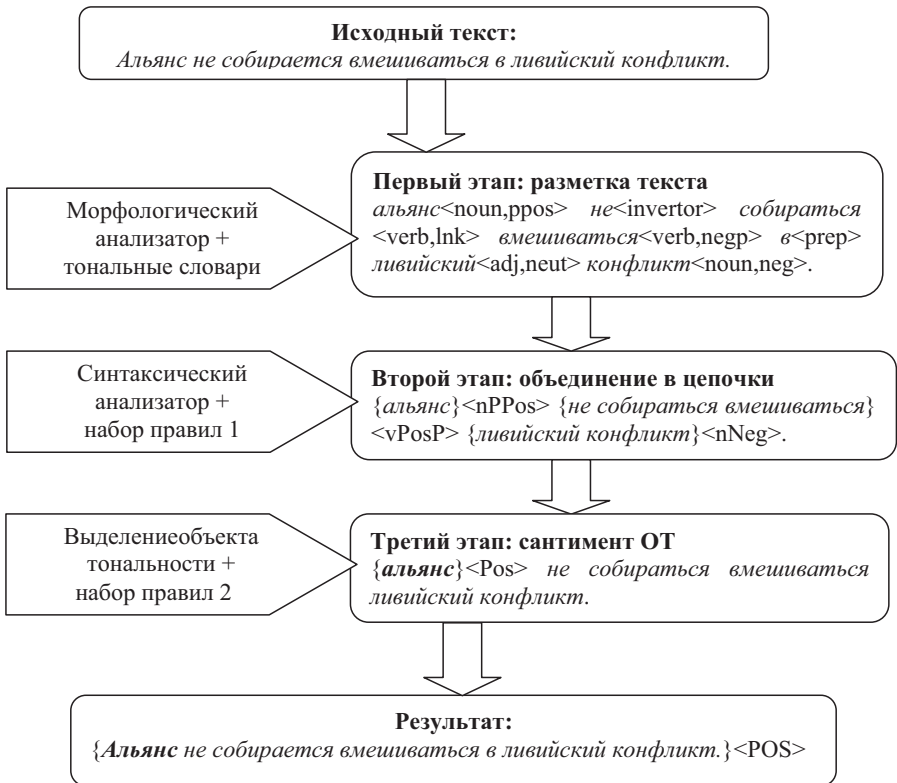
На последнем этапе выделяется объект тональности. Он задается пользователем или определяется автоматически: в каждом предложении ищется так

называемая именованная сущность, например, имя собственное, одушевленное существительное и т. п.

Единица текста, на которой подсчитывается тональность — предикация, и согласно принятому в системе ограничению в каждой предикации тональность считается относительно только одного объекта. Это значит, что в предикации не может быть двух объектов тональности, и при наличии нескольких именованных сущностей исследуемого типа выбирается одна из них.

На основании роли и местоположения объекта тональности ему по определенным правилам приписывается сантимерт, и пользователю предъявляется предложение с выделенным объектом тональности и выявленными тональными цепочками.

Всего было составлено более 20 таких правил приписывания сантимерта объекту тональности.



**Рис. 1.** Последовательность обработки предложения для определения его тональности. Сокращения: noun — существительное, adj — прилагательное, verb — глагол, prep — предлог, invertor — инвертор, pos — позитивный, neg — негативный, negp — чисто негативный, lnk — глагол-связка, ppos — потенциально позитивный, neut — нейтральный, posp — чисто позитивный



Этапы обработки предложения в системе представлены на Рис. 1.

Для тестирования устойчивости системы к добавлениям и изменениям комбинаторных правил и словарей мы создали небольшой тестовый тонально размеченный корпус, охватывающий различные комбинации тональных словоформ и синтаксиса предложений (около 400 предложений). С помощью него мы оцениваем улучшение или ухудшение модуля при каждом значительном изменении системы.

#### 4.4. Оценка результата

В настоящее время методов объективного тестирования систем тональной разметки текстов еще не разработано. Поэтому применяемый в настоящее время нами метод тестирования основывается на периодических субъективных оценках небольших текстовых подборок экспертом. Тестирование проводится один раз в неделю на произвольных текстах СМИ, а именно, используются первые 5–7 новостных текстов с сайта rbc.ru за понедельник или вторник каждой недели, что составляет в среднем по 70 предложений в неделю. Таким образом, за период с января по конец марта 2011 г. система была протестирована на 762 предложениях.

Эксперт получает тональную разметку текстов при помощи системы и затем оценивает, насколько он в каждом конкретном случае согласен или не согласен с результатом. В случае несогласия отмечается тип ошибки: пропуск тональности, неправильный знак тональности (позитив вместо негатива или наоборот), нетональное предложение, размеченное как тональное. Затем на основании этих оценок считается полнота и точность тональной разметки.

Поскольку основной единицей определения тональности является предикация, содержащая не более одного объекта тональности, то полноту и точность тональной разметки также разумно оценивать на основе количества предикаций, на которых тональность сработала правильно или же допустила ошибку какого-либо типа. Таким образом, количество предикаций с верно выделенным объектом тональности и верно определенной тональностью будет соответствовать количеству верных срабатываний системы (А).

Среди ошибок нужно разделять пропуски тональных предложений, содержащих объект тональности (В), и ложные срабатывания — случаи, когда система неправильно определила знак тональности (С) или же выделила как тональное предложение, не содержащее эмоциональной оценки и/или объекта тональности (D).

Эксперт заносит свою оценку тональной разметки, данной системой, в таблицу, строки которой соответствуют предложениям исходного текста, а столбцы — типам срабатываний. Тем самым, в клетках ставится количество выделенных в данном исходном предложении предикаций с верным срабатыванием, с пропусками, с неверным знаком и с «лишней» найденной тональностью (см. Табл. 1).

**Табл. 1.** Фрагмент таблицы с экспертной оценкой результатов работы модуля тональности. Жирным шрифтом выделены найденные модулем объекты тональности, фигурные скобки означают границы тональных предикаций, <POS> — позитив, <NEG> — негатив

№	Предложение	ОК (A)	Пропуск (B)	Знак (C)	Лишнее (D)
42	{В результате взрыва на АЭС «Фукусима-1» поврежден реактор}<POS>.			1	
43	{На четвертом реакторе АЭС «Фукусима-1» в 11: 53 по местному времени (05:53 мск) произошел взрыв водорода}<NEG>, передают японские СМИ.	1			
44	В 11: 14 по местному времени (05:00 мск) в зоне четвертого реактора начался <b>пожар</b> , сообщили в компании-операторе станции Tokyo Electric Power( <b>TEPCO</b> ).		1		

Для подсчёта точности и полноты число предикаций с верными срабатываниями и с ошибками разных типов по всем текстам суммируется, и общая оценка вычисляется по следующим формулам. Число тональных предикаций в тексте составит  $A + B + C$ , а число предикаций, определённых системой как тональные —  $A + C + D$ . Тогда полнота определения тональности будет равняться  $A / (A + B + C)$ , а точность —  $A / (A + C + D)$ .

Система определения тональности постоянно дорабатывается во всех своих частях: словарей, правил, лингвистической базы, программной реализации, поэтому качество определения тональности на текстах, близких к тем, на которых проводится тестирование (новостных СМИ), растёт. Изменение качества определения тональности с января по конец марта 2011 г. представлено на Рис. 2.

Хорошо заметно, что появление в тестовом массиве большого количества не вполне обычных для системы текстов приводит к замедлению роста качества. Так, 14.03 два из пяти текстов в тестовой выборке оказались биржевыми сводками, в которых много специфической лексики, отсюда потеря в полноте.

Можно выделить три класса ошибок, возникающих при определении тональности:

1. Ошибки работы модуля морфологической и синтаксической разметки текста (около 5–7%).
2. Ошибки правил комбинаторики (не более 3%)
3. Ошибки тональных словарей, вызванные их неполнотой и «тональной» омонимией (не более 5%)

Демонстрационная версия модуля тональности доступна по адресу <http://x-file.su/tm/>. Напоминаем, что система рассчитана на работу с грамматически правильными текстами СМИ.

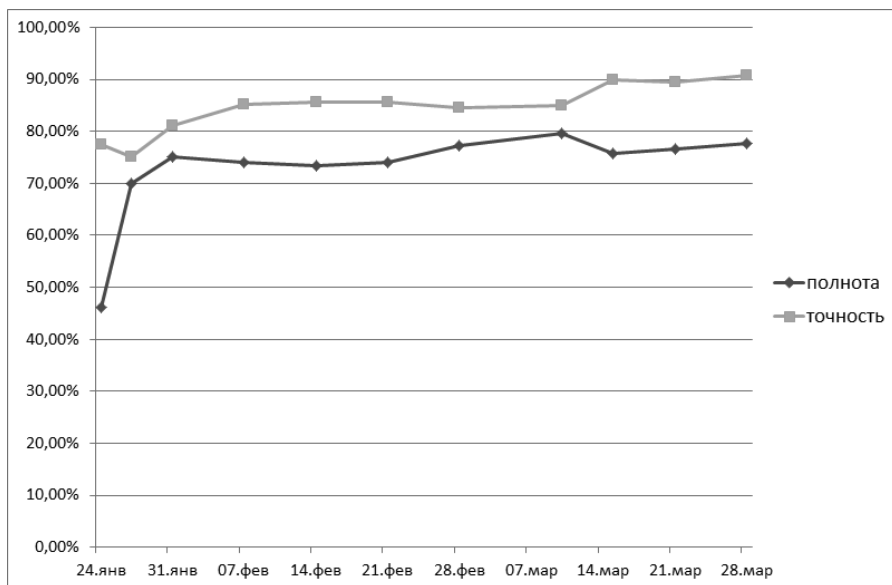


Рис. 2. Изменение качества тональной разметки с января по март 2011 г.

## 5. Заключение

Представленный метод определения тональности относится к так называемому глубокому сентимент-анализу (deep sentiment analysis), основывающемуся на лингвистическом анализе текста на естественном языке (NLP). Как показывают результаты, с помощью этого метода можно достичь достаточно высокой (85–90%) точности на текстах определенной тематики (в нашем случае — новости СМИ). Тем не менее, остается ряд неисправимых ошибок (не учитывая ошибки внешних модулей, такие как ошибки морфологического и синтаксического анализаторов). По нашему мнению, одной из причин такого рода ошибок является ограниченность используемого эмотивного пространства: часть лексики не попадает (или только частично попадает) в наше эмотивное пространство хорошо-плохо плюс сила эмотивности<sup>5</sup>. Определение размерности — открытый исследовательский вопрос, решение которого лежит в области понимания и восприятия информации мозгом человека. Таким образом,

<sup>5</sup> Такое пространство можно назвать полуторамерным — мы не учитывали оппозицию сильный — слабый.

качественное улучшение выбранного нами метода определения тональности нуждается в дальнейших фундаментальных исследованиях не только в области лингвистики, но и в области когнитивных наук, таких как психология, психо- и нейролингвистика.

## Литература

1. *Гаспаров Б. М.* Язык, память, образ. Лингвистика языкового существования // М.: «Новое литературное обозрение», 1996. — 352 с.
2. *Осгуд Ч., Дж. Суси, П. Танненбаум.* Приложение методики семантического дифференциала к исследованиям по эстетике и смежным проблемам // Искусствоведение. Методы точных наук и семиотики / Под ред.: Лотман Ю. М., Петров В. М.. М.: ЛКИ, 2007 г. — С. 278–297.
3. *Соловьев А. Н., Разумихин Д. В., Викторова К. О.* «А ты сам-то как думаешь?» (Об использовании неинформационных функций в моделях речевой коммуникации) // Компьютерная лингвистика и интеллектуальные технологии. Сб. научных статей / Под ред.: Н. И. Лауфер, И. М. Кобозевой, В. П. Селегея. М., Наука, 2003. — С. 653–657.
4. *Якобсон Р. О.* Лингвистика и поэтика // Структурализм: «за» и «против» М. 1975.
5. *Bollen J., Mao H., and Zeng X.-J.* Twitter mood predicts the stock market // Technical Report arXiv:1010.3003, CoRR, 2010. Http: <http://arxiv.org/pdf/1010.3003v1>
6. *Gamon M.* Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis // In Proceedings of the 20th international conference on computational linguistics (COLING 2004), Geneva, Switzerland, August 23–27, 2004 — pp. 841–847.
7. *Konig A. C. & Brill E.* Reducing the human overhead in text categorization // In Proceedings of the 12th ACM SIGKDD conference on knowledge discovery and data mining, Philadelphia, Pennsylvania, USA, August 20–23, 2006 — pp. 598–603.
8. *Nasukawa T. & Yi J.,* Sentiment analysis: capturing favorability using natural language processing // In Proceedings of the 2nd international conference on Knowledge capture, Florida, USA, October 23–25, 2003. — pp. 70–77.
9. *Norcross J. C., Guadagnoli E. & Prochaska J. O.* A visual map of public mementos and conjectures // Journal of Clinical Psychology. v.40, 2006 — P. 1270–1277.
10. *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.
11. *Pang B. & Lee L.* (2005). Seeing stars: exploiting class relationships for sentiment categorization with respect to rating scales // In Proceedings of the 43rd annual meeting of the Association for Computational Linguistics (ACL), University of Michigan, USA, June 25–30, 2005 — pp. 115–124.
12. *Pang B., Lee L., & Vaithyanathan S.* Thumbs up? Sentiment classification using machine learning techniques // In Proceeding of the conference on empirical methods in natural language processing (EMNLP 2002), Philadelphia, PA, USA, July 6–7, 2002 — pp. 79–86.

13. *Prabowo R. and Thelwall M.* Sentiment analysis: A combined approach. *Journal of Informetrics*, 3(2), 2009.
14. *Wilson T., Hoffmann P., Somasundaran S., Kessler J., Wiebe J., Choi Y., Cardie C., Riloff E., Patwardhan S.* (2005) OpinionFinder: A system for subjectivity analysis. *Proceedings HLT/EMNLP, Vancouver (BC)*, 34–35.
15. *Yi J., Nasukawa T., Niblack W. & Bunescu R.* Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques // *In Proceedings of the 3rd IEEE international conference on data mining (ICDM 2003)*, Florida, USA, November 19–22, 2003 — pp. 427–434.