Разработка и реализация алгоритма тоновой классификации коротких текстов

Выполнила – студентка гр.1208 Иванова М.Н.
Научный руководитель – Загорулько Ю.А.
Рубцова Ю.В.

Цель работы: разработка тонового классификатора коротких текстов.

Постановка задачи: разработать алгоритм тоновой классификации коротких текстов.

Этапы работы:

- Исследование предметной области;
- Получение корпуса текстовых сообщений для тренировки тонового классификатора;
- Проведение морфологического анализа текста;
- Проведение анализ SVM и Naïve Bayes
- Выбор и реализация, метода Naïve Bayes
- Проведение экспериментов

Предметная область

Социальные сети, в данной работе используется Twitter.

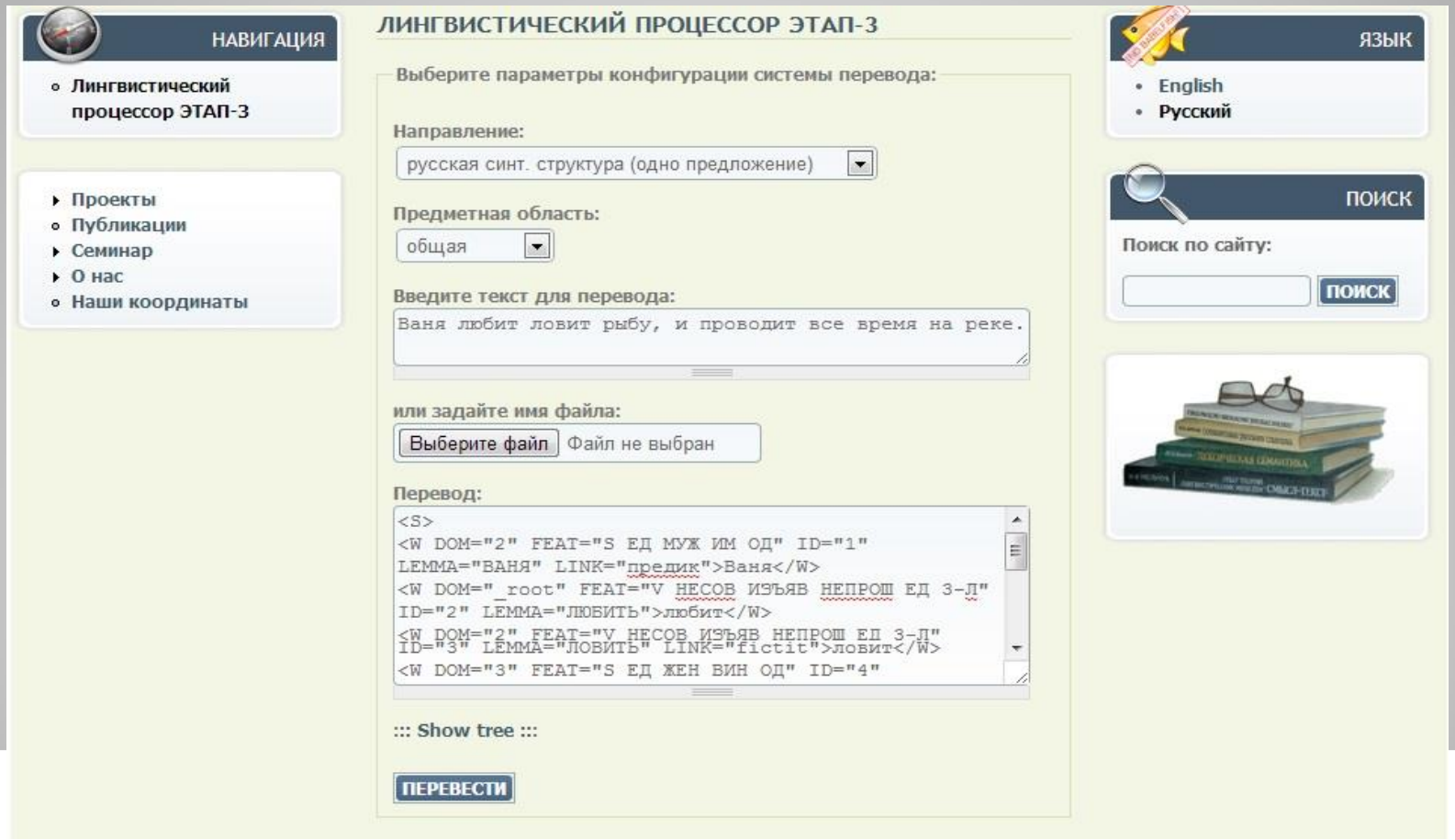
Получение корпуса текстовых сообщений для тренировки тонового классификатора

Текстовые сообщения были получены при помощи API Twitter, основной характеристической величиной являлся «смайл»

Получено три корпуса текста: «Positive», «Negative», «neutral»

Морфологический анализатор корпуса текста.

Анализатор «Лаборатории компьютерной лингвистики» «Института проблем передачи информации им А.А. Харкевича»



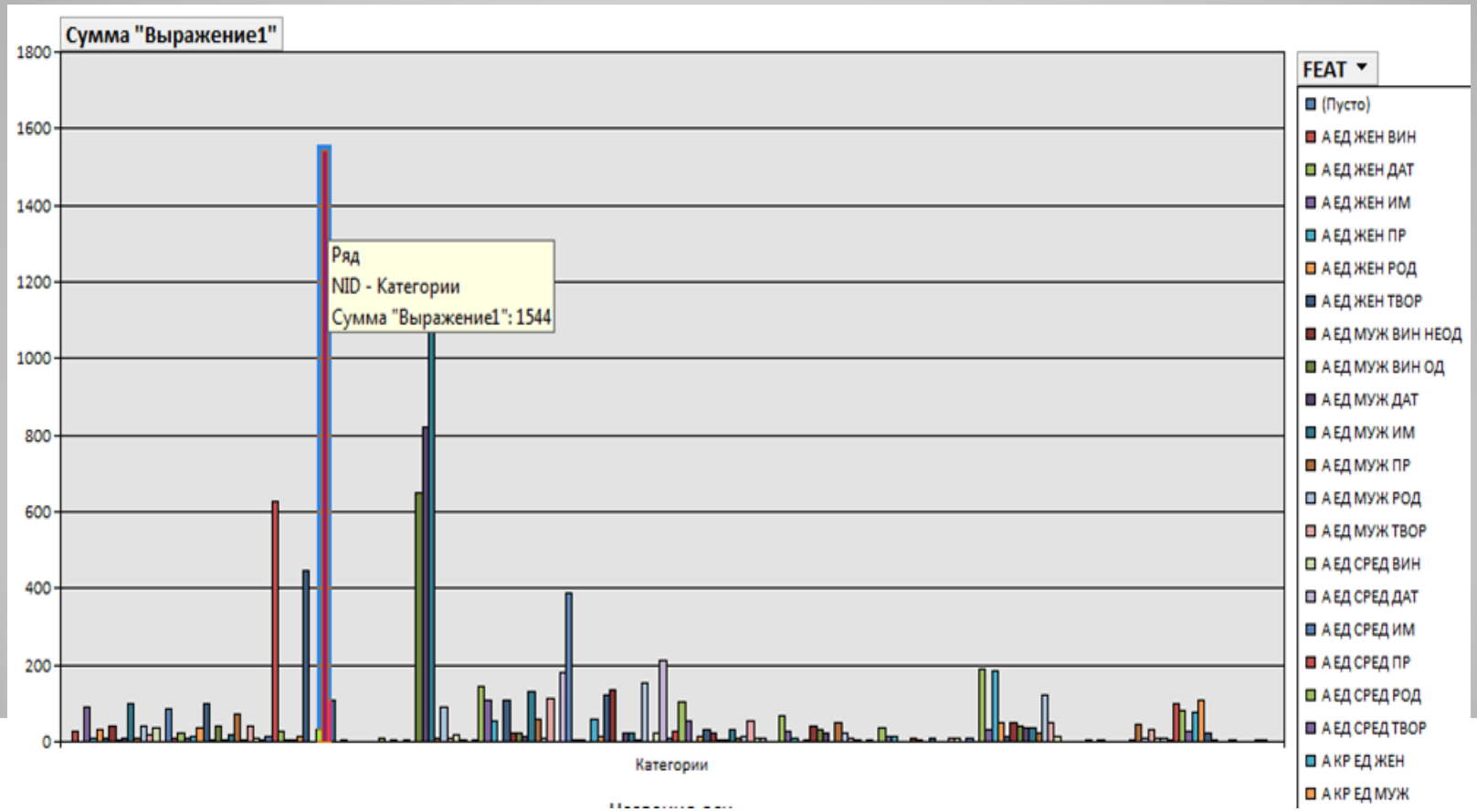
После морфологической обработки ранее полученных коллекций сообщений, было получено три корпуса текста с разбиение слов по морфологическим признакам.

Пример таблицы:

FEAT	LEMMA	LINK	SMS	Код
S ЕД МУЖ ИМ ОД	ПАПА	предик	Папа	23
ADV	КОНЕЧНО	вводн	конечно	24
S ЕД МУЖ ИМ ОД	ХОЗЯИН	предик	хозяин	25
PR	В	разъяснит	в	26
S ЕД МУЖ ПР НЕОД	ДОМ	предл	доме	27
CONJ	НО	сент-соч	НО	28
S ЕД ЖЕН ИМ ОД	МАМА	предик	МАМА	29
S ЕД МУЖ ИМ ОД	ХОЗЯИН	соч-союзн	хозяин	30
S ЕД МУЖ РОД ОД	ПАПА	1-компл	папы	31

Оценка полученных данных.

Результат обработки корпуса текстов «positive»



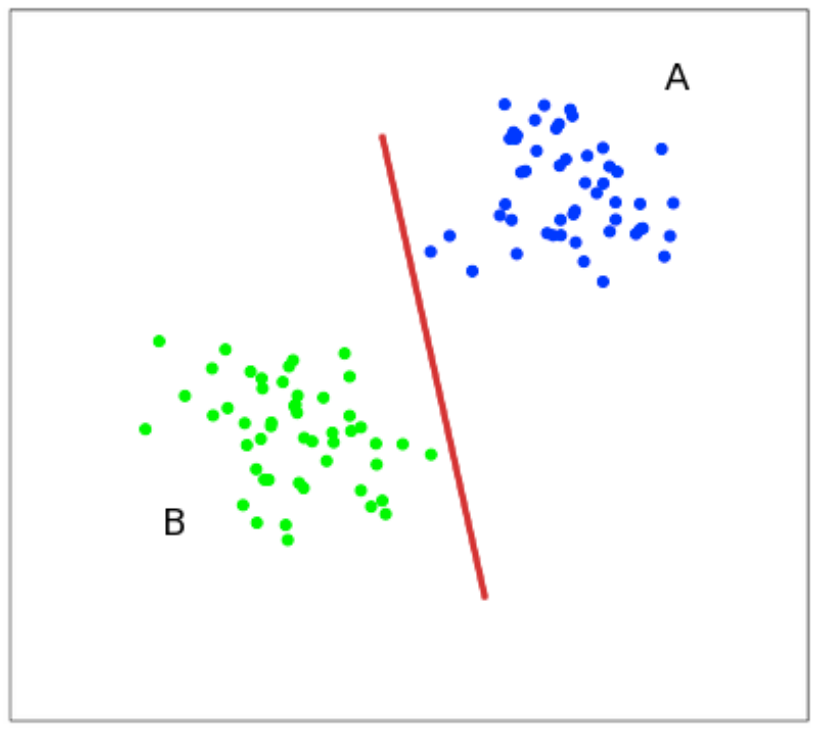
Результаты, полученные после морфологического анализа.

Результатом являются три коллекции текстов, разобранных по морфологическим признакам.

Анализ алгоритмов

- Метод опорных векторов
- Метод Naïve Bayes

Метод опорных векторов



Naïve Bayes

Наи́вный ба́йесовский классифика́тор — простой вероятностный классификатор, основанный на применении теорема Баеса со строгими (наивными) предположениями о независимости.

Реализация метода Naïve Bayes

Коллекция текстов для обучения классификатора.

@relation "myrelation"
@attribute txt string
@attribute classatr {positive,negative,neutral}
@data
‘txt’, positive // одно текстовое сообщение принадлежащее классу «positive»
‘txt’, negative
‘txt’, neutral

В данном файле содержится 2/3 (2,600) текстовых сообщений всего корпуса, где текстовых сообщений принадлежащих к корпусам текстов positive, negative, neutral одинаковое количество.

Коллекция для тренировки содержит 1/3 (1,200) всего корпуса текста.

ЭКСПЕРИМЕНТЫ

1. Работа с текстовыми сообщениями помеченными эмоциями.

Результат:

Правильные экземпляры 13 81.25 %
Не правильные экземпляры 3 18.75 %
Количество экземпляров 16
81\

1. Работа с текстовыми коллекциями после морфологического анализа.

Результат:

Correctly Classified Instances 15395 48.2133 %
Incorrectly Classified Instances 16536 51.7867 %
Total Number of Instances 31931

Вероятностные ошибки

Идея заключается в том что мы притворяемся как будто видели каждое слово на один раз больше, то есть прибавляем единицу к частоте каждого слова.

Логически данный подход смещает оценку вероятностей в сторону менее вероятных исходов. Таким образом, слова которые мы не видели на этапе обучения модели получают пусть маленькую, но все же не нулевую вероятность.

Вывод

Эксперименты показали, что при работе с полными текстовыми коллекциями, важно использовать наиболее точный корпус текстов, при обучении.

Рекомендуется проверять обучающий текст проверять на стоп слова.

Результат проделанной работы

В ходе данной квалификационной работы были выполнены следующие задачи:

- Изучены предметная область
- Проведен анализ метода опорных векторов(SVM) и метод Naïve Bayes.
- Выбран и реализован, метод Naïve Bayes
- Проведены эксперименты

Дальнейшее развитие данной работы может быть связано с исследованием тональности текстов художественной литературы, так как данная работа посвящена коротким текстовым сообщения из социальных сетей.

Список литературы

1. Read, J (2005). Using Emoticons to Reduce Dependency in Machine Learning Techniques for Sentiment Classification. In: Proceedings of the Student Research Workshop at the 2005 Annual Meeting of the Association for Computational Linguistics. Ann Arbor, Michigan, pp.43-48.
2. Логашенко И.Б. (2011)Современные методы обработки экспериментальных данных,
3. Н.В. Лукашевич (louk_nat@mail.ru) НИВЦ МГУ, Москва, И.И. Четверкин (ilia2010@yandex.ru) Факультет вычислительной математики и кибернетики МГУ, Москва АВТОМАТИЧЕСКАЯ КЛАССИФИКАЦИЯ ОТЗЫВОВ НА ОСНОВЕ ОЦЕНОЧНЫХ СЛОВ,
4. Bo Pang and Lillian Lee Department of Computer Science Cornell University, Ithaca, NY 14853 USA {pabo,llee}@cs.cornell.edu; Shivakumar Vaithyanathan IBM Almaden Research Center 650 Harry Rd. San Jose, CA 95120 USA shiv@almaden.ibm.comSentiment Classification using Machine Learning Techniques,
5. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis, (2008)
6. Peter D. Turney Institute for nformation Technology National Research Council of Canada Ottawa, Ontario, Canada, K1A 0R6 (2002) Canada.Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification