УДК 004.6-024.11:004.738.5

АНАЛИТИЧЕСКИЙ КОМПЛЕКС АНАЛИЗА ДАННЫХ ИЗ ОТКРЫТЫХ ИНТЕРНЕТ ИСТОЧНИКОВ

И.И. Пилецкий
Доцент кафедры информатики БГУИР, кандидат технических наук, доцент

В.А. Прытков
Декан факультета компьютерных систем и сетей БГУИР, кандидат технических наук, доцент

Н.А. Волорова
Заведующая кафедрой информатики БГУИР, кандидат технических наук, доцент

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: ianmenski@gmail.com

Аннотация. В этой статье приводится описание прототипа инструмента мониторинга открытых интернет источников с целью выявления бренда в некоторой предметной области, определения тематик публикаций, оценки тональности публикаций, оценки соотношения с другими брендами. Данная статья описывает принятые решении для нескольких задач в области аналитики Big Data в рамках одного проекта. Здесь под термином бренд понимается: объект, субъект, событие, явление.

Ключевые слова: интернет-СМИ, Big data, мониторинг, анализ, машинное обучение, Apache Spark, обработка естественного языка, тональность

Введение. В настоящее время анализу неструктурированных текстовых данных, получаемых из социальных сетей посвящено много работ. Эти работы носят как экспериментальный, пробный характер, так и для принятия решения. Как правило, решения принимаются на основании применения некоторого метода анализа данных. Здесь под анализом данных, могут рассматриваться события, явления, товары и услуги компаний, отдельные персоны и т. п. Результатами анализа смогут воспользоваться компании для создания систем поддержки пользователей, социологи для проведения анализа общественного мнения, организаторы мероприятий для получения отклика участников, знаменитости для отслеживания репутации в сети, правительство для контроля настроений в обществе и др.

Несмотря на наличие различных систем для анализа данных из Интернет-СМИ, данная тематика является актуальной, ибо в настоящее время, в социальный сетях, можно найти разнообразную информацию о почти любых явлениях в мире, о деятельности организаций и людей. Среди наиболее популярных систем можно выделить - GATE [1] и UIMA [2]. Европейская комиссия разработала собственный инструмент для мониторинга и анализа Интернет-СМИ, Европейский медиа-монитор **(**EMM) [3], который, среди прочих функций, позволяет пользователям ознакомится с основными новостями на многих языках. На фоне достаточно широкого спектра существующих технологических решений для текстового анализа, следует подчеркнуть, что эти инструменты существуют либо как «разработка открытых конкретных проектов", или как специализированные запатентованные инструменты фиксированной функциональности, что не позволяет использовать данные решения для создания универсального комплексного проекта с расширяемой и модифицируемой библиотеки аналитических модулей, которые могут применяться для анализа некоторых данных различными методами.

В настоящее время в ряде областей деятельности основное внимание уделяется индивидууму, конкретному человеку, явлению и оценки влияния на конкретный объект. То есть, акцент анализа переноситься с общей оценки события на оценку действий конкретного индивидуума и принятия решений на влияние на его действия.

В данной статье рассматривается проект создания «Аналитического комплекса анализа данных из открытых интернет источников», позволяющего получить информацию популярности некоторого бренда (компании, объекта, товара, события, явления и возможно некоторого человека) на основе анализа существующих данных из электронных газет, журналов, статей, социальных сетей с помощью различных аналитических методов и оценить отношение в Интернет-СМИ к этому явлению. Значимость проекта состоит в создании наращиваемого многомодульного комплекса, позволяющего анализировать большие данные из интернет источников различными аналитическими методами и выборе для применения наиболее подходящих для данного явления.

Технология построения и функционирования Аналитического комплекса. Аналитический комплекс из состоит компонент (см. рис. 1): сбора данных, фильтрации данных и составление «мешка слов» из N-грамм (векторизации), хранилища данных, библиотеки аналитических модулей, подготовки выдачи результата, клиентского модуля.

При необходимости набор модулей компонент может быть расширен, а некоторые модули заменены новыми.

Сами компоненты состоят из функциональных модулей, на рис. 1 приведена логическая схема взаимодействия компонент и модулей.

Аналитический комплекс выполняет свои функции в несколько этапов:

−компонент сбора данных − выполняет целевое сканирование выбранных социальных сетей, новостных порталов, сайтов и помещает полученные документы в специальном виде в базу данных документов (в хранилище);

−компонент обработки данных выполняет фильтрацию исходных данных (токенизацию, лемматизацию, стемминг, удаление стоп слов, перевод к нижнему регистру), построения «мешка слов» и векторизацию исследуемых наборов документов («мешок слов» в специальном виде храниться в хранилище);

−компонент библиотека аналитических модулей, содержит набор модулей, которые осуществляют обработку данных, полученных из интернет источников с целью поиска упоминаний о брендах, определения их тональности и формирования аналитических данных для передачи клиентскому модулю, а также содержит управляющие и служебные модули. Данный компонент анализа данных содержит модули, реализующие как самообучающиеся NLP алгоритмы, так и NLP алгоритмы с учителем, позволяющие выполнить не только семантический анализ данных (например, отношение к бренду), но и построить тематическую модель коллекции текстовых документов (например, кластеризацию и классификацию документов). Компонент библиотека аналитических модулей может содержать модули: SVM, VSM, Neural networks, Naive Bayes, LDA, PLSA, Logistic Regression, Decision tree. В настоящее время апробированы векторные и вероятностные модули. Применение комбинированного подхода к оценке некоторого явления позволяет наиболее достоверно его оценить, выполнить ретроспективный анализ прошедшего события, идентифицировать новое события и принять правильные решения;

−компонент подготовки выдачи результата – подготавливает информацию в виде отчетов для пользователей аналитического комплекса;

−графический интерфейс для взаимодействия с пользователем системы обеспечивает клиентский модуль, который принимает от пользователя запросы, определяет тип запроса и вызывает нужный модуль как для их выполнения, так и для выдачи результата. Так, для выдачи результата передает запрос компоненту анализа для подготовки и выдачи аналитических отчетов;

–компонент хранилища данных – содержит данные из социальных сетей, предварительно обработанные и размеченные данные, необходимые для построения классификатора, «мешок слов», информацию о брендах, а также служебную информацию, необходимую для работы других модулей системы.

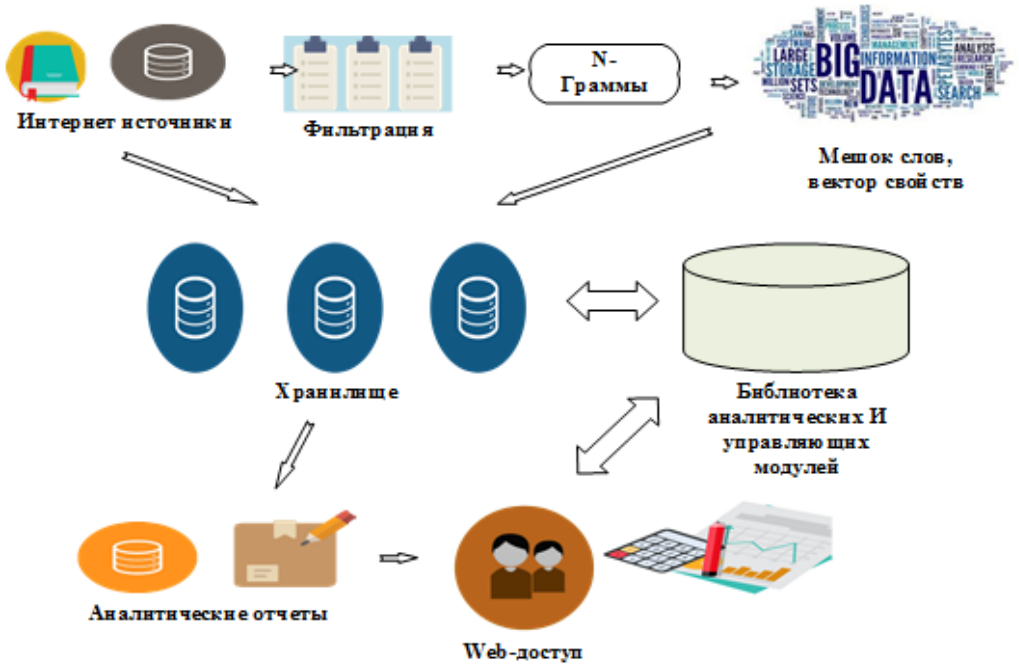


Рисунок 1. Логическая схема взаимодействия компонент

Технологическая платформа. В качестве технологической платформы использовались решения на платформе Open Source Apache Hadoop и облачная платформа IBM Bluemix для проведения экспериментальных решений. В частности, использовались: Apache Ambari – система управления, мониторинга и администрирования Hadoop кластера, Apache Hadoop - с файловой системой HDFS и менеджером ресурсов YARN, Apache Spark - фреймворк для обработки данных с MLib, Apache Hbase - распределенная база данных, Apache Kafka - фреймворк для обработки потоковых данных, Apache Nutch и др. для чтения данных с Web-страниц.

Для экспериментальных работ использовались IBM Bluemix cloud platform (сервисы Apache Spark) и IBM Watson Analytics for Social Media (сервис Natural Language Understanding, позволяющий выполнить фильтрацию данных), библиотека BigARTM для тематического моделирования больших коллекций текстовых документов, библиотеки ***SciPy и*** NumPy для языка программирования Python.

Примеры реализации.

Пример 1. Качество результатов обучения и работа алгоритма SVM (*accuracy, precision, recall, f-measure)* [4].

Обучение на корпусе твитов русского языка: 114291 документ положительной тональности и 111020 отрицательной и корпусе твитов английского языка 562024 документа положительной тональности и 534912 отрицательной (твиты).

Для построения «мешка слов» опробованы все 11 весовых схем: Binary, Tf, tf-idf, tf-rf, tf-or, tf-ig, tf-mi, Prob-Based, tf-icf, ICF-Based, ΔTF-IDF.

Качество классификации оценивалось по всем характеристикам: Accuracy, Precision, Recall и F-Measure по всем весовым схемам.

Опытным путём установлено, что на корпусе русскоязычных документов наилучшие были получены для схемы tf-idf: 0.91, 0.89, 0.93, 0.91, а для корпуса документов на английском

языке 0.93, 0.95, 0.94, 0.94 соответственно.

Применение. Коллекцией текстовых сообщений, собиравшейся с социальной сети Twitter с 01.02.2016 по 31.01.2017, результаты см. рис. 2 и рис. 3.

Таблица 1

Бренды для поиска

Название бренда	Список ключевых выражений
MANCHESTER UNITED	MANCHESTER UNITED, MAN UNITED, MUFC
MANCHESTER CITY	MANCHESTER CITY, MAN CITY, MCFC
ARSENAL	ARSENAL, GUNNERS
TOTTENHAM	TOTTENHAM, THFC
PREMIER LEAGUE	EPL, PREMIER LEAGUE, PREMIERLEAGUE, PL
CHAMPIONS LEAGUE	CHAMPIONS LEAGUE, CHAMPIONSLEAGUE, UCL
FA CUP	FA CUP, FACUP
EFL CUP	LEAGUE CUP, EFLCUP, EFL
EUROPA LEAGUE	EUROPA LEAGUE, UEL

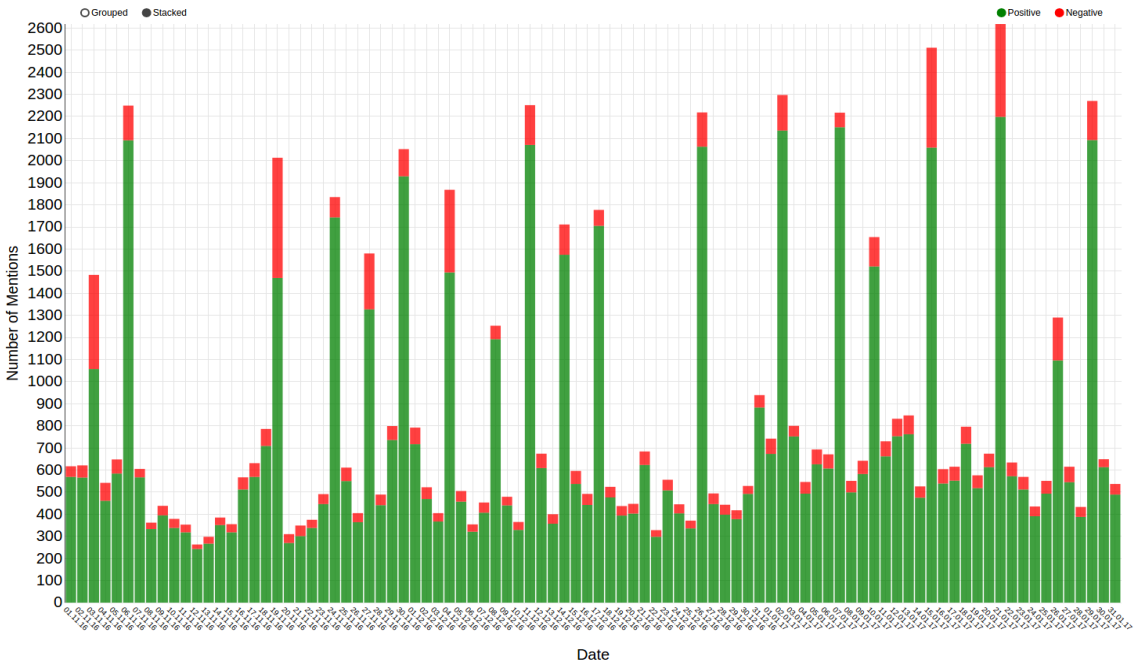


Рисунок 2. Динамика тональности упоминаний бренда Manchester United

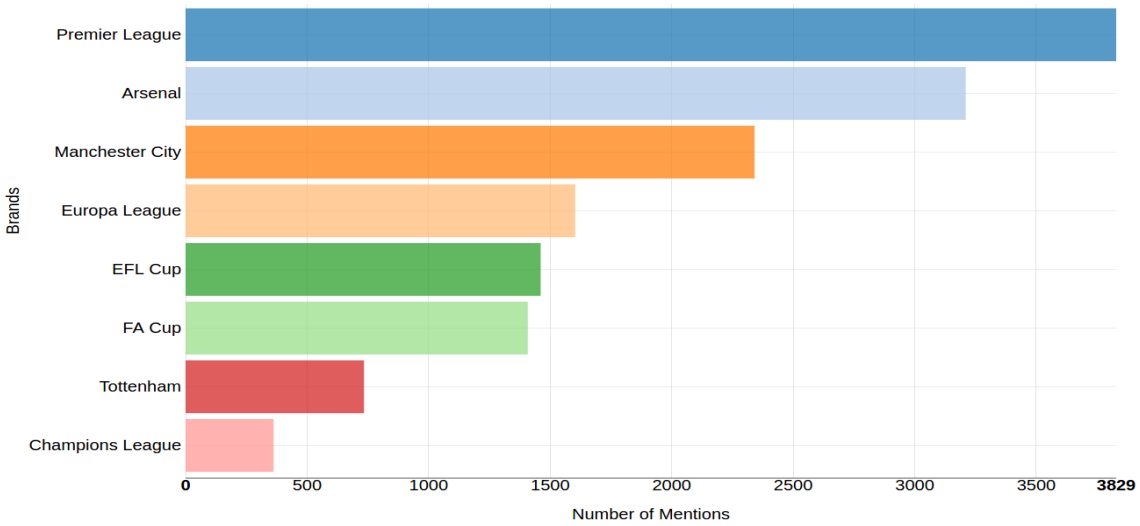


Рисунок 3. Бренды, которые упоминались с Manchester United

Пример 2. Реализация алгоритмов кластеризации LDA, PLSA [5].

Обучение без учителя. Тематический анализ текста Интернет-новостей, Интернет-СМИ (на пример: *Трамп, выборы, Клинтон, политика, США, дебаты и т.д.*), оценка сантиментов новостей из газет (положительная, нейтральная, отрицательная).

Фильтрация текстов (нормализации текстов) с помощью сервиса IBM Watson Analytics for Social Media (сервис Natural Language Understanding).

EM-алгоритм использовался для обучения, и оценка качества алгоритмов определялась, как перплексия (perplexity) $Perplexity = \exp\left(-\frac{1}{2}\sum_{d\in D}\sum_{\omega\in D} n_{d\omega}\, ln\big(p(\omega|d)\big)\right)$. Чем меньше эта величина, тем лучше модель предсказывает появление слов ω в документах d корпуса D $(d \in \mathrm{D})$. Сравнение двух алгоритмов LDA, PLSA показало, что алгоритм PLSA более точный по сравнению с алгоритмом LDA.

Ниже приведены результаты анализа статей с сайтов foxnews.com и dw.com.

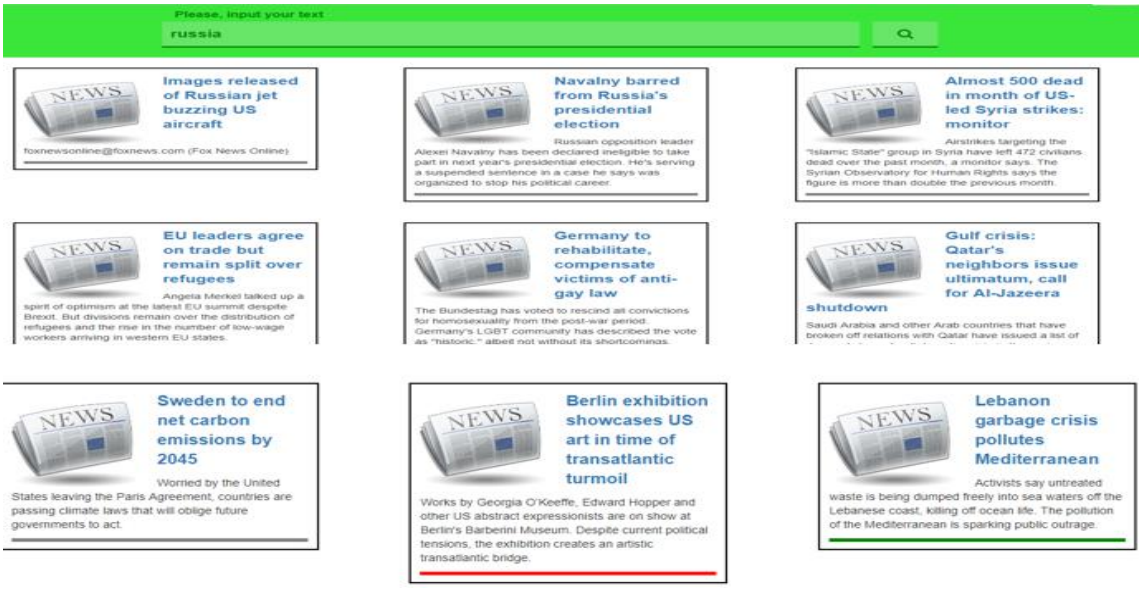
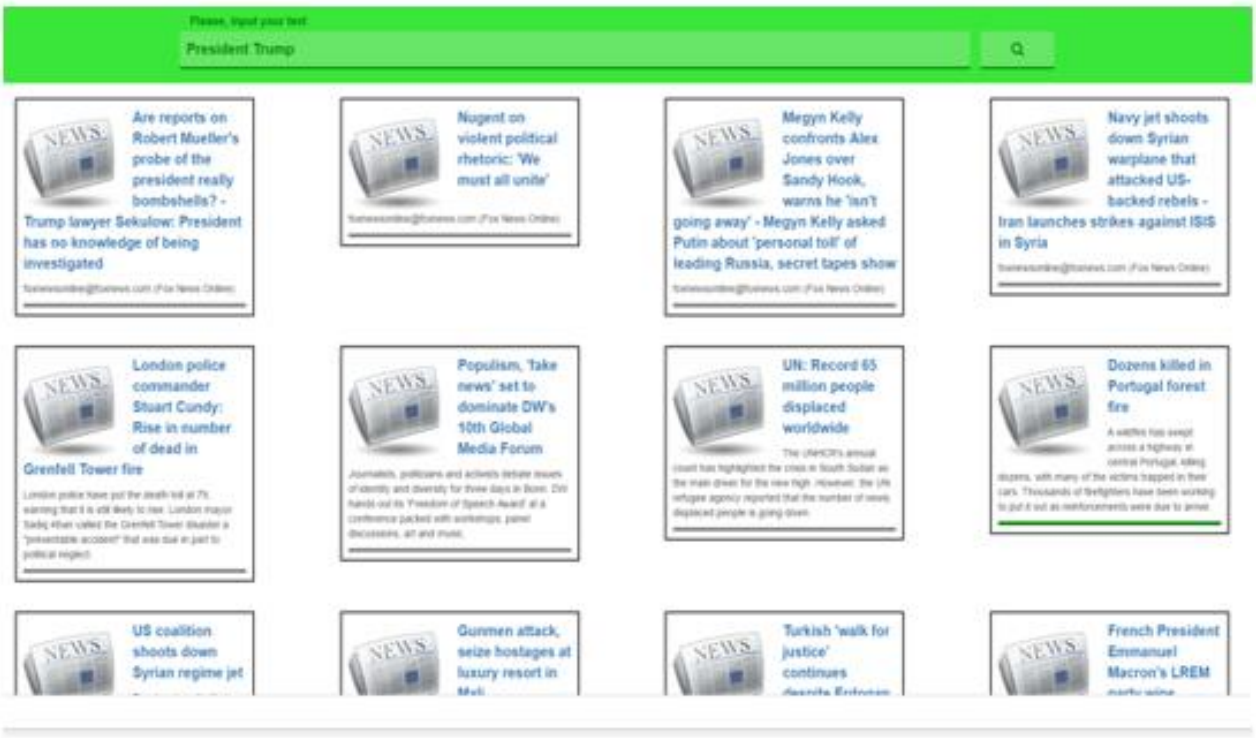


Рисунок 4. Результаты по запросу «*russia*»



Рисуок 5. Результаты по запросу «*President Trump*»

Пример 3. Реализация алгоритмов классификации тем PLSA

Обучение без учителя. Тематический анализ текста и в качестве похожести тем (для классификации) используется косинусная мера.

Данные с сайтов: Onliner.by, Tut.by, Kyky.org.

Фильтрация текстов (нормализация текстов) и получение «мешка слов» по описанной технологии Аналитического комплекса.

Классификация. Тематический анализ с использованием EM-алгоритма позволяет выявить n самых важных тем во всём тексте, на основании словаря соответствий конкретных, заранее определенных тем, получаем темы для определения конкретных классов. Достоверность определения правильных тем в зависимости от параметров приблизительно равно 82%.

Список литературы

[1]. Books about GATE [Электронный ресурс] / Режим доступа: https://gate.ac.uk/books.html. Дата доступа: 12.03.2018.

[2]. Welcome to the Apache UIMA project [Электронный ресурс] / Режим доступа: https://uima.apache.org. Дата доступа: 12.03.2018.

[3]. Europe Media Monitor [Электронный ресурс] / Режим доступа: http://emm.newsbrief.eu/overview.html. Дата доступа: 12.03.2018.

[4]. Романов А.А., Пилецкий И. И. Классификация тональности текстовых документов с помощью метода опорных векторов. Компьютерные системы и сети: материалы 53-й научной конференции аспирантов, магистрантов и студентов. – Минск: БГУИР, 2017 -06 мая 2017.

[5]. Чугаинов К. В., Пилецкий И. И. Методы тематической кластеризации новостных статей. Научно-практические исследования №2 (ISSN 2541-9528) – Омск: Дельта, – 2017 с. 295 – 298.

ANALYTICAL COMPLEX ANALYSIS OF DATA FROM OPEN INTERNET SOURCES

I.I. PILETSKI, PhD
Associate Professor of Informatics Department of the BSUIR

V.A. PRYTKOV, PhD
Dean of the Faculty of Computer Systems and Networks of the BSUIR, Associate Professor

N.A. VOLARAVA, PhD
Head of the Chair of Informatics of the BSUIR, Associate Professor

Belarusian State University of Informatics and Radioelectronics, Republic of Belarus
E-mail: ianmenski@gmail.com

Abstract. This article describes the prototype of the monitoring tool for open Internet sources in order to identify the brand in a certain subject area, identify topics for publications, evaluate the tonality of publications, assess the relationship with other brands. This article describes the decisions taken for several tasks in the field of Big Data analytics within a single project. Here, by the term brand is understood: object, subject, event and phenomenon.

Key words: internet media, Big data, monitoring, analysis, Machine Learning, Apache Spark, Natural Language Processing, sentiment, tonality.