Scientific Bulletin of the „Petru Maior" University of Tîrgu Mureş
Vol. 9 (XXVI) no. 1, 2012, ISSN 2285 – 438X (Online), ISSN–L 1841 – 9267

SENTIMENT ANALYSIS FRAMEWORK ORGANIZATION BASED ON TWITTER CORPUS DATA

Adela BERES

Department of Mathematics and Informatics,
Faculty of Sciences and Letters
„Petru Maior" University of Tîrgu Mureş, Romania
Nicolae Iorga Street, no. 1, 540088, Mureş County

adela.beres@upm.ro

ABSTRACT

Since its inception in 2006, Twitter has gathered millions of users. They post daily tweets about news, events or conversations. These tweets express their opinion about the topic they are discussing. Twitter is a large database of content that can be semantically exploited to extract opinions and based on these opinions to classify the users. This paper presents the organization of a sentiment analysis framework based on Twitter corpus data, including crawling tweets and opinion mining of the tweets, making it easy for its users to create portfolios of trustful Twitter accounts.

Keywords: social media, sentiment analysis framework, Twitter corpus data, Twitter crawling module, Twitter accounts

1. Introduction

For companies operating in the financial market it is crucial to take the right decisions at the right moment. This means being able to maximize the profit based on relevant information that they hold or obtain.

This information comes from employees, technical analysts or other external sources. Today, however, is quite hard to know which persons are reliable and which are truly visionaries in this field.

Social Media has developed enormously in the recent years. Social Media is a series of web and mobile applications that are created from user-generated content. Types of Social Media applications are: collaborative projects (Wikipedia), community sharing (YouTube), blogs and micro blogs (Twitter), social networks (Facebook), virtual worlds (Second Life) etc.

Since it was started in 2006, Twitter has gathered millions of users. Messages posted to Twitter are called tweets. They have the following characteristics:

- the messages contain maximum 140 characters
- if the message is addressed to a particular user it contains the @ symbol followed by the username of that user
- when addressing a community, area, topic or group this is marked with an hashtag #
- retweets are marked with RT

Usually a tweet should be concise, clear, addressed and connected.

Twitter represents an enormous database that can be exploited in terms of semantic analysis. Almost every person has a Twitter account and constantly posts messages regarding news, events or daily chatter.

In this context an enormous amount of information is stored on the web. Human processing capacity fails to exploit this source. The data is not linked in order to make correlations. As Tim Berners-Lee said: "we are drowning in information but starving for knowledge".

This paper presents the organization of a sentiment analysis framework of an application on Trust level based on Twitter corpus data. Twitter users are classified based on the opinions expressed in their tweets.

Based on this classification, portfolios of users can be created. These portfolios contain only those users whose information, decisions and capabilities are confident and worth taking into account when making financial decisions.

Two directions of analysis can be taken into account when classifying users:

- the moment when the user tweeted
- the sentiment or opinion expressed by the user

2. Sentiment Analysis

The emergence of the web has changed the way people communicate with each other, their information structure, documentation and exchange of views. Social networking, online communities, blogs, Social Media in general has resulted in millions of texts and documents in which the authors express their feelings and opinions about products, events, people. All this information has become a huge database for IR (information retrieval) and sentiment analysis.

This data, however, is usually unstructured or very different depending on who wrote it or the field for which it was written, making the task of analyzing it very difficult even for a human. If that person is unfamiliar with the language and concepts expressed the task becomes even more complicated.

Sentiment analysis is often compared with IR (information retrieval), but is a more difficult task because we have to analyze not only the text itself but also the semantic meaning, determine the opinion, whether it is negative or positive and discover the attributes included in that opinion.

Sentiment analysis is the computational study of the views, opinions and emotions expressed by people for certain entities, events and their attributes. It attracted much attention from both academic and industrial forums because of the research issues raised and of the wide range of applicability. (Bing, 2010)

Sentiment analysis operates with the following concepts:

- The object and its attributes
- Opinion holder
- Opinion or orientation

The object is the entity for which the opinion is being expressed. It can be anything from persons to events, organizations, etc. These objects are described by a set of attributes such as size, color, location, etc.

Opinion holder is the person expressing the opinion or feeling. This may be the author of the text or another person.

The opinion expressed about an object is a positive or negative view that the opinion holder has about the object.

Attributes of an object can be selected and presented in several ways: unigrams, n-grams, lemma, negation, opinion words or adjectives. Attributes and their values are usually grouped into vectors. (Boiy et al., 2007)

Techniques used in sentiment analysis fall into two categories:

- Based on a lexicon
- Machine learning

Each of them is based on a conceptual modeling of the environment, i.e. fall in the space of model classification with regard to description and ontology. (Rădoiu, 2008)

Lexicon-based techniques treat text as a bag-of-words, or a set of unrelated words. Depending on the text certain words may be pozitive or negative. In order to extract the opinion mainly adjectives or adverbs are used. Usually they are used in combination with a noun or verb to create a context. These combinations are then searched in AltaVista or WordNet to see if they are "close" to a word with positive polarity or "close" to a word with negative polarity. Based on the results a classification is made of words that express positive, negative or neutral feelings.

Machine learning techniques use two different methods: supervised and unsupervised. The Naive-Bayes, Maximum Entropy and SVM (Support Vector Machines) are supervised methods.

Naive-Bayes classifiers use Bayes rule which states that class attributes are independent of each other. It uses the distribution of the number of attributes found in the document.

Maximum Entropy classifiers use entropy models to choose from a variety of models. Models behave like class rules in the document, and the most suitable model is the one that satisfies the most rules.

SVM (Support Vector Machines) is not a probabilistic technique. Hyperplanes are used to build the training sets. The model is a representation of the opinion words as points in space, mapped so that the words of different categories are divided by a clear gap that is as wide as possible.

The most important problem in the case of entity extraction and sentiment analysis is the fact that these techniques were mostly designed for documents that have a well-defined structure. However, social awareness streams are short with very little structure in terms of syntax. Therefore a corresponding research challenge is to establish how these methods must be modified or completely redesigned to work well also in the case of such micro-documents. (Stan, 2011)

3. Crawling Twitter

Microblogs, such like Twitter, allow users, after registering with a username and password, to post short messages called tweets. Most microblogs allow users to select also the group to which they want to send those messages.

Users of these online communities use microblogging sites to distribute different types of information. A recent review of Twitter revealed the use of microblogging for:

- Daily chatter – for example what one is currently doing
- Conversations – for example tweets to the followers group
- Exchange of information – for example links to websites
- News – for example titles of articles

Despite the diversity of users of this simple communication channel, it was observed that they usually fall into two categories:

- Users who post about themselves

- Users who share information (Bollen et al., 2011)

In both cases, the tweet may contain information about the mood and feelings of its author.

Sentiment analysis on Twitter corpus data is quite difficult because tweets are very short, only 140 characters and can be posted from anywhere (computer, phone) thus it's highly probable to make spelling mistakes. The users usually express their feelings using emoticons, for example ☺ for a happy mood or ☹ for a sad mood. They also use acronyms, such as gr8t instead of "great" or lol for "laughing out loud". Another feature is that users can add additional letters to emphasize a feeling, for example "cooooool". Most tweets contain hashtags and links to sites or references to other users making their analysis even more difficult.

Twitter doesn't expose all of its tweets for free and to anybody. It has three access levels:

- Spritzer
- Gardenhose
- Firehose

The Spritzer access level is all free and exposes about 1% of all the public statuses of Twitter. The Gardenhose level exposes about 10% of all public tweets but requires case-by-case approval by Twitter. The most important level is the Firehose which is not free and the user can get access to all the public tweets for an amount of money which Twitter calculates based on how much Twitter thinks the user can pay.

For crawling Twitter exposes the following APIs:

- Search API
- REST API
- Streaming API

The most simple and accessible one is the Search API which is a dedicated API for running searches against the most recent tweets. The tweets that can be crawled are between 6-9 days old. The search is anonymous and it's restricted by the complexity of the query and not by the number of requests per day.

The REST API is, as its name suggests, a REST service to crawl the Twitter posts. Requests can be anonymous or authenticated. In case of authenticated requests OAuth protocol is used. The user gets a ConsumerKey, ConsumerSecret, OAuthToken and an AccessToken which are used when making requests. Based on these he can make up to 350 requests per hour. If he constantly exceeds this limit he will be black-listed and further requests will be denied. In case of the anonymous requests the rate limit is made on the IP address from which the requests are coming from.

The Streaming API is the most powerful API for crawling Twitter. Connecting to the streaming API requires keeping a persistent HTTP connection open. This means that the user doesn't need to worry any more about rate limits. However to many connection attempts can cause the IP to be banned. There are mainly three streams to which the user can connect: public streams (public data on Twitter), user streams (single user view of Twitter) and site streams (multiple users' views of Twitter).

The data requested from Twitter is provided in json format and includes information about the author of the tweet, tweet content, date, geolocation, entities, urls or hashtags.

4. Sentiment Analysis Framework Organization

The main modules that we propose for a sentiment analysis framework based on Twitter corpus data are:

- Database for storing the tweets
- Crawler Twitter
- Semantic Processor
- Web application

The database is a simple SQL database with the following table structure:

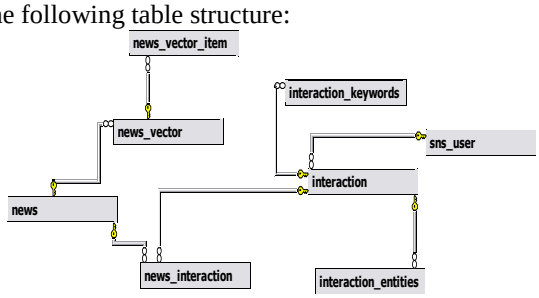

Figure 1 Database structure

The Interaction table represents the tweets from Twitter. It contains the date, content and author of the tweet. The sns_user table contains all the users from Twitter from which we crawled posts. The interaction_keywords and interaction_entities tables contain sentiment analysis data for the existing tweets in the database. This data consists of: keyword/entity count, relevance, score and sentiment.

The News and related tables are a representation of a discussion topic on Twitter. The news is the topic itself, news_vector contains the vectors for that news and the news_vector_items are the actual hashtags, keywords or entities used for crawling Twitter for tweets from that specific topic.

The organization that we propose for the sentiment analysis framework based on Twitter corpus data is:

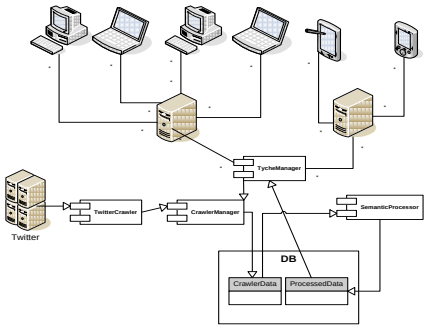

Figure 2 Sentiment Analysis framework organization

The Twitter Crawler module is responsible for querying Twitter for relevant tweets from the financial domain. This module can be a Windows service running permanently. The crawling is done in two ways:

- Crawl tweets from a specific user
- Crawl tweets for a specific topic

Using the Twitter REST API and the OAuth tokens requests to Twitter can be made within a 350 rate limit per hour. For more requests per hour multiple Twitter accounts with corresponding OAuth tokens can be used.

The data crawled from Twitter is then passed to the CrawlerManager which stores it in the database in the corresponding table.

From here the Semantic Processor module takes each interaction entity and processes it to extract the sentiment. The sentiment analysis is done for:

- Keywords
- Entities

Each tweet is analyzed and the keywords and entities found are extracted. Then for each keyword and entity the sentiment is calculated. For the sentiment also a score is given, not only a distinction on positive/negative/neutral.

The Semantic Processor module then stores the data into separate tables for keywords and entities. This module can be a web service easily accessible also for other third-party consumers.

The tweets data and metadata extracted from Twitter and also the semantically processed data is shown to the end user through a web application. Here the user can see the latest tweets, stats about Twitter users, comparisons between topics and other useful information.

5. Conclusions

Twitter, with millions of users posting messages every day, has become an enormous database of content. This data can be exploited semantically using sentiment analysis methods.

Sentiment analysis on Twitter corpus data can be a difficult task because the tweets are short and can contain spelling errors, acronyms, emoticons, hashtags or urls.

We propose an organization of a sentiment analysis framework based on Twitter corpus data which consists of a Twitter crawling module, a semantic processor module and a web application to display relevant information to the end users.

The advantages of this organization are the fact that it is robust and decoupled, each module can perform separately and can be accessed by third-party consumers. These modules can be hosted on separate servers having one common point – the database.

Based on the crawled data and on the sentiment information provided by the framework a user can create its own portfolios of trusted Twitter accounts.

References

[1] Alec Go, Lei Huang, Richa Bhayani, *„Twitter Sentiment Analysis”*, Association for Computational Linguistics, 2009

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, Rebecca Passonneau, *„Sentiment Analysis of Twitter Data”*, Proceeding LSM '11 Proceedings of the Workshop on Languages in Social Media, SUA, 2011

[3] Bing Liu, *„Sentiment Analysis: A Multi-Faceted Problem”*, IEEE Intelligent Systems, 2010

[4] Bing Liu, *„Sentiment Analysis and Subjectivity. Handbook of Natural Language Processing”*, Second Edition, 2010

[5] Changhua Yang, Kevin Hsin-Yih Lin, Hsin-Hsi Chen, *„Building Emoticon Lexicon from Weblog Corpora”*, Proceedings of the ACL 2007 Demo and Poster Sessions, Prague, June 2007

[6] Dumitru Rădoiu, “*Virtual organizations conc eptual modeling*”, Studia univ. Babes Bolyai, I nformatică, vol. LIII, no. 1, 2008

[7] Erik Boiy, Pieter Hens, Koen Deschacht, Marie-Francine Moens, *„Automatic Sentiment Analysis in On-line Text”*, Proceedings ELPUB2007 Conference on Electronic Publishing, Vienna, Austria, June 2007

[8] Johan Bollen, Alberto Pepe, Huina Mao, *„Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena”*, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM 2011), Barcelona, Spain, 17-21 July 2011

[9] Johan Bollen, Huina Mao, Xiaojun Zeng, *„Twitter mood predicts the stock market”*, Journal of Computational Science, 2011

[10] Johann Stan, *“A Semantic Framework for Social Search”,* Universite Jean Monnet, Saint-Etienne, 2011

[11] John Domingue, Dieter Fensel, James A. Hendler, *„Handbook of Semantic Web Technologies”*, Springer, 2011

[12] Jonathon Read, *„Using Emoticons to reduce Dependency in Machine Learning Techniques for Sentiment Classification”*, Proceedings of the ACL Student Research Workshop, USA, June 2005

[13] Luciano Barbosa, Junlan Feng, *„Robust Sentiment Detection on Twitter from Biased and Noisy Data”*, Coling 2010: Poster Volume, pages 36-44, Beijing, August 2010

[14] Matthew A. Russel, “*Mining the social Web*”, O’Reilly, 2011

[15] Matthew A. Russel, “*21 Recipes for Mining Twitter*”, O’Reilly, 2011