A MODEL OF EXTRACTING PATTERNS IN SOCIAL NETWORK DATA USING TOPIC MODELLING, SENTIMENT ANALYSIS AND GRAPH DATABASES

Assane Wade[1] and Giovanna Di MarzoSerugendo[2]

Centre Universitaire d'Informatique
University of Geneva, Geneva, Switzerland

ABSTRACT

Social networks analysis studies the interactions among users when using social media. The content provided by social media is composed of essentially two parts: a network structure of users' links (e.g. followers, friends, etc.) and actual data content exchanged among users (e.g. text, multimedia). Topic modeling and sentiment analysis are two techniques that help extracting meaningful information from large or multiple portions of the text: identifying the topic discussed in a text, and providing a value characterizing an opinion respectively. This extracted information can then be combined to the network structure of users' links for further tasks as predictive analytics, pattern recognition, etc. In this paper we propose a method based on graph databases, topic modelling and sentiment analysis to facilitate pattern extraction within social media texts. We applied our model to Twitter datasets, and were able to extract a series of opinion patterns.

KEYWORDS

Topic modelling, Sentiment analysis, Neo4j, Opinion mining, Twitter, Graph database, pattern.

1. INTRODUCTION

The increasing availability of data sources in recent years has been accompanied by dramatic progress in machine learning theories and algorithms and their application to many domains such as computer vision, speech recognition, natural language processing and predictive analytic, making the data analytic area a prominent and important field of research and exploitation. On the one hand, structured data, gathered by companies as a result of their day-to-day operations, has been widely exploited with Business Intelligence (BI) techniques and tools. Results help decision makers by providing a clear understanding of current situation at hand.

On the other hand, unstructured data (e.g. texts, images, blogs, tweets, etc.) are usually harder to localize (they must be gathered from external sources such a social media, or web blogs) and to mine with classical BI techniques. In fact, a major trend of unstructured data analytics is the use

David C. Wyld et al. (Eds) : CCSEIT, AIAP, DMDB, ICBB, CNSA - 2017
pp. 75– 84, 2017. © CS & IT-CSCP 2017 DOI : 10.5121/csit.2017.70608

of Social Networking Analysis (SNA) theories and methods. This analysis concerns both underlying network structure of users' links (followers, friends, etc) and the actual data exchanged inside these networks.

Exploitation of the structure of the network is usually based on graphs theory, while extraction information from data resulting from the interactions among users of the network is based on a combination of data analytics and text mining. With the increasing informational size of social networks, those data sources have become an important source of informal data related to the activities and environments of the companies, or useful for companies to understand trends or opinions.

Sentiment analysis "also called opinion mining , is the field of study that analyses people's opinions, sentiments, evaluations, appraisals, attitudes, and emotions towards entities such as products, services, organizations, individuals, issues, events, topics, and their attributes"[1]. The main application of sentiment analysis is to study what is said about a service or a product. When people want to purchase a good or service they can search what the other customers think about it.

This make the companies more attentive of what is said about their brand of products. It gives an idea of the acceptance of what they propose. Other applications can be the study of the opinion about a social or political activity, about an event. These studies can be biased or not precise enough when they are applied to social network. Indeed, social network content and discussions are freely flowing, not necessarily classified into themes; for example, a user can make a political post in a discussion about medicine. Topic modelling studies this issue.

Topic modelling studies the classification of documents by topic or theme. This activity is relevant when dealing with a large amount of data. The extraction of data from a social network is based on a keyword search, which are not put in context. We can therefore be redirected to any content which contains our keyword. For example, when we search the keyword "Apple", we are directed to content linked to the fruit as well as to the well-known company. For better accuracy, we apply topic modelling methods to classify this content and use the context we want to study.

We propose a model, based on topic modeling, sentiment analysis and graph databases, that exploits both the network structure of users' links and actual content data from social media. We apply the proposed model to Twitter in order to identify opinion patterns.

2. SENTIMENT ANALYSIS IN A NUTSHELL

Sentiment Analysis is firstly a natural language processing task at many levels of granularity. The first application can be found at the document level classification task[2, 3], later it has been done at the sentence level[4, 5] and recently at the phrase level [6, 7]. "Sentiment analysis or opinion mining is the computational study of opinions, sentiments and emotions expressed in text" [6].This opinion extraction from text is a complex task in the context of social network, this is particularly more challenging with micro-blogs such as tweets and blog reviews. The main challenge is decoding the text [8, 9]. This is because of the non-formal writing style inside these media. One reason can be the limitation of the number of characters in some social media.

2.1 Document level sentiment analysis

This method focuses on the entire document. For each **d ∈ D,** a set of documents, this method computes the value of the sentiment for d. The features are not taken in account or are just assumed as the object; in addition, the opinion of a document is considered as expressed by one opinion-holder. An example is [6], where this approach was used at a document-level to classify movie reviews into two classes, positive and negative. A drawback of the method is we can have multiple opinion holders for long texts. For example, a blogger can refer the opinion of other people for the sake of a comparison.

2.2 Sentence level sentiment analysis

In sentence level analysis, we have two tasks:

- The subjectivity classification: determines if the sentence is an objective or subjective sentence, by searching for opinionated sentences [10,11];
- Sentiment classification: computes the polarity of the subjective sentences.

Sentences have a single opinion expressed by one opinion holder [13,14].

2.3 Feature based sentiment analysis

The assumption in feature based sentiment analysis is that the overall opinion does not mean complete acceptance of every aspect of an object [12, 15]. One can positively mark a product without liking all the aspect of the product. In our example, one can say that the computer is a good one but simultaneously find the screen too large. Feature based sentiment analysis tries to capture this phenomenon. We have the overall polarity on an object and a polarity of each of its features. The following tasks are performed:

- Identify the features of an object.
- Determine the polarity of the features.

2.4 Sentiment analysis approaches

a) Machine Learning Approach

We identify the two following methods.

Unsupervised Learning: these methods classify sentiment by assuming some classification rules. Turney [7] used this method to find the sentiment on a review by using some syntactic rules.

Supervised Learning: These methods use labeled data to compute the polarity([15], [16]). The model is trained to exploit the training data and then tested with a training set. Supervised Learning is the most used machine learning method in sentiment analysis. The first challenge is to

label the data automatically. The algorithms like naive Bayes (NB),maximum entropy(ME), and support vector machines (SVM) are often used in sentiment analysis.

b) Lexicon-Based Approach

Lexicon-Based approach uses a dictionary of opinionated words. Each word of the dictionary has a polarity. We notice three approaches:

- Manual approach: this method needs a lot of time to be built. Compared to automatic methods, this method is seldom used.

- Dictionary based approach: this dictionary is built automatically. The system starts with a pre-defined list of opinionated words (called seeds) and new words are added when they appear. The list is augmented using synonyms of the words in the seed from online dictionary like WordNet

- Corpus-based approach: this method is more powerful than the others, as it includes the syntactic rules and co-occurrence rules pattern. Another element in this method is the contextualization of a word. This is very important because a word can have a positive or negative polarity based on the domain.

The above different approaches propose a model of capturing the opinion patterns based on database technologies, and an assumed pre-identified topic.

3. MATERIALS AND METHODS

We first discuss here the system architecture we propose and then explain each step of the process we designed.

3.1 System architecture

The process we designed combines topic modelling, sentiment analysis and a graph database storage. Our system (Figure 1) is composed of three layers. These levels communicate in a bidirectional manner. We describe here each level of the architecture:

- Sources: We consider only tweets with a text content (we do not consider multimedia tweets) The whole content data is extracted from the source and stored in the storage (RDBMS)

- Storage: this layer includes all the kind of storage we use during all the process. In our case we use a relational database for content data and a graph database for storing the network structure of users' links.

- Application: this covers all the programs used during the process: topic modeling, sentiment analysis, and opinion pattern extraction.

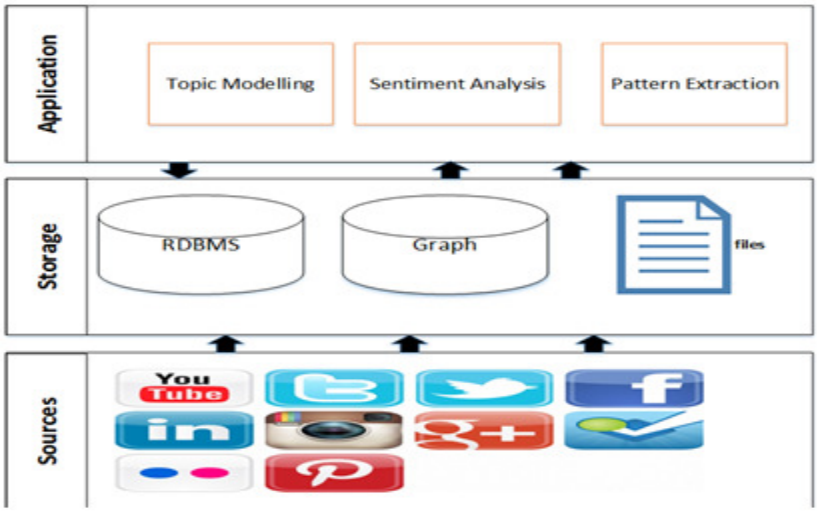

Figure 1.System architecture

The process starts by crawling data from Twitter followed by a cleaning phase. We then perform topic modelling using LDA (see Section 3.2 below). For each dataset we do a first run of LDA to identify the most relevant topic we want to study (we call it "Main Topic"), ie, the topic that gathers more tweets. We then run a second time LDA on this topic to extract sub-topics. Once we performed the sub-topic modelling we need to prepare the sentiment analysis by finding the remaining information included in the MySQL database. To run topic modelling, we just extracted the text from the database. To continue the analysis we extracted the user_name, the screen_name and other metadata linked to each tweet. The sentiment analysis is then done. For each tweet we have the Id of the user, his name, the time it was posted, the text of the tweet, the sentiment polarity. We store all this in a file. We created another file containing the relation between users in the file (the followers' graph). These two files are then imported into the graph database. From this database we then extract the pattern of opinion in the network for the different sub-topics identified earlier.

3.2 Latent Dirichlet Allocation (LDA)

Topic modeling techniques are designed to discover statistically latent topics inside a collection of documents. The first known method is probabilistic latent semantic indexing (pLSI) sometimes called Latent Semantic Analysis (LSA) [17]. In this method, each word in a document is a sample from a mixture model where topics are represented as the multinomial random variables and documents as a mixture of topics. The approach makes three assumptions:

- The semantic information can be derived from a word-document co-occurrence matrix
- This dimensionality reduction is an essential part of this derivation
- And the words and documents can be represented as points in Euclidean space.

LDA is an unsupervised machine learning technique used to identify random topic information in large document collections. It is based on a "bag of words" approach.[18]

In this approach, each document is seen as a vector of word counts. Each document is represented as a probability distribution over some topics, while each topic is represented as a probability distribution over a number of words. The generative process used by LDA in a collection for each document is:

- For each document, pick a topic from its distribution over topics.
- Sample a word from the distribution over the words associated with the chosen topic.
- The process is repeated for all the words in the document.

We applied LDA on all tweets, identified a relevant topic, and focused the rest of our study on the tweets that relate to that topic. We additionally manually evaluated the tweets to ensure they are pertinent. From here we run a second time LDA to extract sub-topics to have a more fine-grained list of topics.

3.3 Sentiment Analysis

Vader (Valence Aware Dictionary for Sentiment Reasoning) is based on a powerful and extendible lexicon for sentiment analysis [19]. We have decided to use it because of its social network oriented capabilities. The reason is that the sentiment analysis methods are not all appropriated to social network text, in particular micro-blogging such as tweets. With VADER we need no training of the model because it has already been trained and tested. Another advantage of Vader is the calculation of the strength of the sentiment. The strength goes from extremely negative to extremely positive. In our case we have defined 5 polarity types: extremely negative, negative, neutral, positive and extremely positive. We performed the sentiment analysis with Vader Sentiment implementation in Python. For each tweet we extract its sentiment. We extract the strength of the sentiment which is called compound. The compound is a number between –1 and+1. We then assign to each tweet its polarity types based on the compound value. Table 1 shows the polarity type and the lower and upper bound of each polarity. The definition of the name of the polarity will help us querying the graph database.

Table 1. Classification of sentiment analysis

Polarity Type	**Compound**	
	Lower Bound	**Upper bound**
ExtremelyPositive	> 0.5	1
Positive	0	0.5
Neutral	0	0
Negative	<0	-0.5
Extremely Negative	< -0.5	-1

3.4 Neo4j

Neo4j is a native graph databases management system with NoSQL capabilities written in JAVA. The system is very strong and easy to deploy in a personal computer. It gives a web base interface for visualizing the graph.

Figure 2 represents the database model. We have three types of nodes:

- User: this node represent a user of our dataset: its properties are a unique id and a unique screen_name. We have reflexive links between users that represents a user following another user. We name the relation Follow.

- Tweet: represents a tweet published by a user. That's why we define a relation between node user and node Tweet (Publish_A). The Tweet node contains the id of the tweet, the time it was posted and the text of the tweet.

- PolarityType: is a node that stores the polarity of a tweet contained in node Tweet. The two nodes are connected by a relationship called Has_Polarity. The properties of node Polarity are the value of the sentiment (Compound) and the polarity type (extremely positive, positive, neutral, negative and extremely negative).

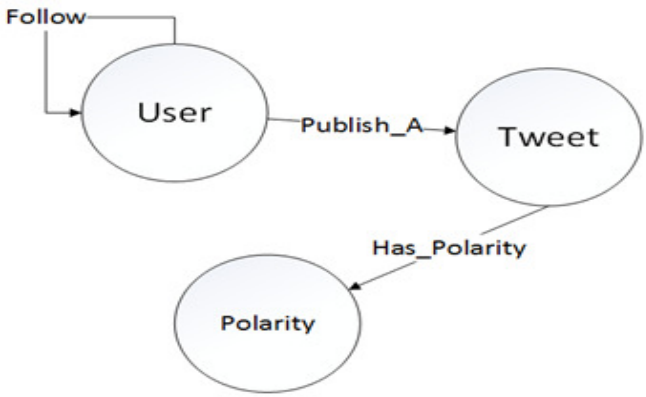

Figure 2. Graph database model

To define and manipulateNeo4j databases we use a declarative graph query language called Cypher. This query language is based on Pattern matching. It allows finding a node or a group of node based on specific conditions. It also allows implementing some graph theory algorithms like the shortest path discovery, calculate the degree of a node etc. Cypher and SQL are very similar in the way they build queries.

4. IMPLEMENTATION RESULTS

We discuss here the application of our model to a Twitter dataset in order to extract opinion change patterns.

4.1 Dataset

Our data set is a mixed dataset obtained from the Twitter API using a series of keywords related to different topics (such as movie, Obama, Tesla).This dataset contains 600,000 tweets crawled during one month period in April 2016.

4.2 Results

Figure 1shows the key elements of each step of the process. From the 600,000 tweets on our dataset, we had 460,743 tweets after the cleaning process. Then we applied topic modelling on this last number of tweets and produced 46,254tweets for the topic we considered as having the most relevant content (the "main topic"). This set of tweets has been processed a second time in order to extract the sub-topics (10 sub-topics). The sentiment analysis of the 46,254 tweets constituting the chosen main topic shows that the majority of opinionated tweets in that topic were positive (20%) or extremely positive (24%).

Table 2 shows the structure of the results we obtained from this process.

Table 2. Results Summary

Topic Modelling	Number of tweet:600,000 Number of tweet of the chosen topic:46,254 Number of subtopics: 10 Number of Users: 25,624
Sentiment polarities	Extremely positive: 24 % Positive: 20% Neutral: 50% Negative: 2% Extremely negative: 4%
Follower Graph	Number of users in the follower graph: 235 Number of connections: 420

The graph database stores the information of Figure 2, as a result of the process we discussed above. For a given user and his followee, we can now extract opinion pattern of that user for a given topic using the Cypher query language. Figure 3 shows the tweets of a user (in red) following three other users. The opinions of the different users are very heterogeneous in the example. But the positive sentiments are more present than negative ones.

In the same manner we can extract other opinion changes in the database for all users and topics. This provides an overview of the trend of sentiment in the database. We can also organize the results according to the network structure of users' links (e.g. number of followees, followers…).

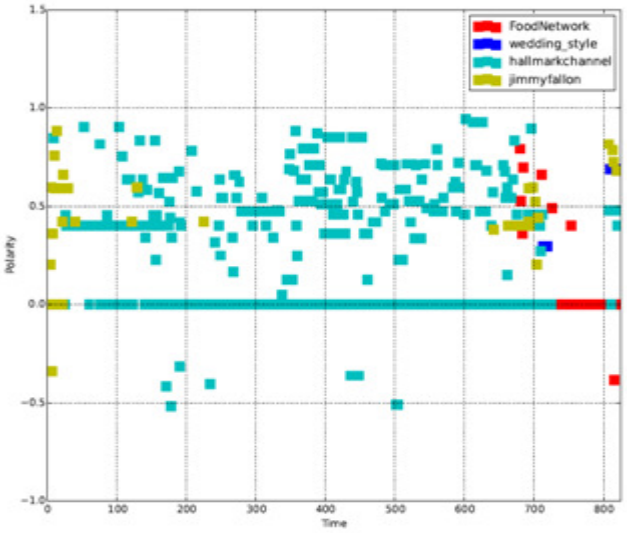

Figure 3. Example of pattern of opinion

5. CONCLUSIONS

We proposed a model for extracting patterns of opinion using topic modelling, sentiment analysis and the great opportunities provided by graph databases. We applied the model to a Twitter dataset. The model allowed us to extract opinion by combining the network of user's followers links with content data. Future works will: (1) study recurrent opinion changes across users and topics in order to identify opinion change patterns; (2)understand the dynamics of opinions change within social networks; extracting other patterns (recurrent discussions, volume of discussions…). The work can also be extended to community detection [20]. This can lead to a method of detecting influencers withincommunities.

REFERENCES

[1] B. Pang et L. Lee. Opinion Mining and Sentiment Analysis.Found. Trends Inf. Retr., vol. 2, pp. 1-135,jan 2008.

[2] T. Nasukawa et J. Yi. Sentiment analysis: Capturing favorability using natural language processing.in Proceedings of the 2nd international conference on Knowledge capture, 2003.

[3] H. Kanayama et T. Nasukawa. Fully automatic lexicon expansion for domain-oriented sentiment analysis.in Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 2006.

[4] M. Gamon, S. Basu, D. Belenko, D. Fisher, M. Hurst et A. C. König. BLEWS: Using Blogs to Provide Context for News Articles..inICWSM, 2008.

[5] T. Joachims. Making large scale SVM learning practical.1999.

[6] B. Pang et L. Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts.in Proceedings of the 42nd annual meeting on Association for Computational Linguistics, 2004.

[7] P. D. Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews.inProceedings of the 40th annual meeting on association for computational linguistics, 2002.

[8] M. Hu et B. Liu. Mining and summarizing customer reviews.InProceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004.

[9] S.-M. Kim et E. Hovy. Determining the sentiment of opinions.InProceedings of the 20th international conference on Computational Linguistics, 2004.

[10] J. M. Wiebe, R. F. Bruce et T. P. OHara. Development and Use of a Gold-standard Data Set for Subjectivity Classifications.InProceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, Stroudsburg, 1999.

[11] A. Agarwal, B. Xie, I. Vovsha, O. Rambow et R. Passonneau. Sentiment Analysis of Twitter Data.InProceedings of the Workshop on Languages in Social Media, Stroudsburg, 2011.

[12] N. Agarwal et H. Liu. Modeling and data mining in blogosphere.Synthesis lectures on data mining and knowledge discovery, vol. 1, pp. 1-109, 2009.

[13] H. Yu et V. Hatzivassiloglou. Towards Answering Opinion Questions: Separating Facts from Opinions and Identifying the Polarity of Opinion Sentences.InProceedings of the 2003 Conference on Empirical Methods in Natural Language Processing, Stroudsburg, 2003.

[14] V. Hatzivassiloglou et J. M. Wiebe. Effects of Adjective Orientation and Gradability on Sentence Subjectivity.InProceedings of the 18th Conference on Computational Linguistics - Volume 1, Stroudsburg, 2000.

[15] L.-W. Ku, H.-W. Ho et H.-H. Chen. Novel Relationship Discovery Using Opinions Mined from the Web.InProceedings of the 21st National Conference on Artificial Intelligence - Volume 2, Boston, 2006.

[16] V. M. K. Peddinti et P. Chintalapoodi. Domain Adaptation in Sentiment Analysis of Twitter.InProceedings of the 5th AAAI Conference on Analyzing Microtext, 2011.

[17] T. Hofmann. Probabilistic Latent Semantic Indexing.InProceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, New York, NY, USA, 1999.

[18] D. M. Blei, A. Y. Ng et M. I. Jordan. Latent Dirichlet Allocation.J. Mach. Learn. Res., vol. 3, pp. 993-1022, #mar# 2003.

[19] C. J. Hutto et E. Gilbert. VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text.InProceedings of the Eighth International Conference on Weblogs and Social Media, {ICWSM} 2014, Ann Arbor, Michigan, USA, June 1-4, 2014., 2014.

[20] L. Chunshan, C. William K, Y. Yunming, Z. Xiaofeng, C. Dian-Hui et L. Xin. The Author-Topic-Community model for author interest profiling and community discovery.Knowl. Inf. Syst, vol. 44, n° %12, pp. 359-383, 2015.