Sentiment analytics using RapidMiner

Introduction

"If you torture the data long enough, it will confess."
- Ronald Coase.

The developments of the web have transformed the way businesses function today. Companies are leveraging the web to connect with customers in multiple ways. The websites, mobile apps and social networks are facilitating the companies in listening to their existing and prospective customers, build relationship with them and is enabling cross/up selling. While the web is providing many opportunities, companies are confronted with many difficulties as well. One of the major challenges is pertaining to harvesting actionable insights from the huge amount of data that is generated by the growing number of touch points facilitated by the web to connect with the customers. Humongous amount of data is being generated every day, but ironically most organizations are data rich but information poor (Minhaj, 2016).

Brands have been criticized, loved, hated, recommended and patronized on the web. A plethora of platforms like the discussion forums, blogs, social networks are used by the customers to discuss the brands and give explicit/implicit and positive/negative feedback about their experience of using the products/services. The negative sentiments expressed on the web not only affect the current customers, it also affects future prospects. Hence, it becomes imperative for the companies to closely monitor their online

reputation and handle their customer sentiments, especially the negative sentiments in an efficient and timely manner.

What is Sentiment Analysis?

Opinions are central to almost all human activities because they are key influencers of our behaviours. Whenever we need to make a decision, we want to know others' opinions (Sentiment Analysis and Opinion Mining, 2016). The businesses and organizations always want to find consumer or public opinions about their products and services. Individual consumers also want to know the opinions of existing users of a product before purchasing it. Generally, individuals ask their friends or relatives when they are in need of an opinion. On the other hand, organisation conventionally have been conducting surveys, opinion polls when they need public or consumer opinion. However, in the contemporary business world, with the explosive growth of web platforms like forums, blogs, micro-blogs and social networking sites, individuals and organizations are increasingly using the content in these media for decision making. Organization are no longer limited to conducting surveys and opinion polls to gather public opinion because there is a profusion of such content publicly available. However, understanding the intent from that content remains a formidable task. It is in this context that the domain of text mining, particularly the automatic mining of opinions from text sources is emerging as a widely used approach. This process of extracting and understanding the opinions from text sources is commonly referred both by the scholarly community and the industry as Sentiment Analysis.

Sentiment analysis is the process of computationally identifying and categorizing opinions expressed in a piece of text, especially in order to determine whether the writer's attitude towards a particular topic, product, etc. is positive, negative, or neutral (Oxford Dictionaries, 2016)

The approach

Sentiment analysis or opinion mining basically deals with the study

of analysing people's opinions, sentiments, evaluations, attitudes, and emotions from written language. This domain has emerged as a one of the most active research areas in Natural Language Processing and Text Mining. A plethora of approaches have been devised towards sentiment analysis by different researchers. While the different approaches differ in their degree of accuracy of determining the right sentiment, the premise of most approaches entail classifying the opinions into neutral or polar, that is positive or negative.

The task of classifying the text into neutral, positive or negative can be performed at different levels of the texts :

Document
Sentence
Entity – feature/aspect

The sources of data that are commonly used by companies to analyse their customer sentiments include:

Emails
Reviews
Feedback data
Tweets
Blog entries

The most important indicators of sentiments are sentiment words, also called opinion words. These are words that are commonly used to express positive or negative sentiments. For example, good, wonderful and amazing are positive sentiment words. Likewise, bad, poor and terrible are negative sentiment words.

Sample indicators of Positive Sentiments	Sample indicators of Negative Sentiments
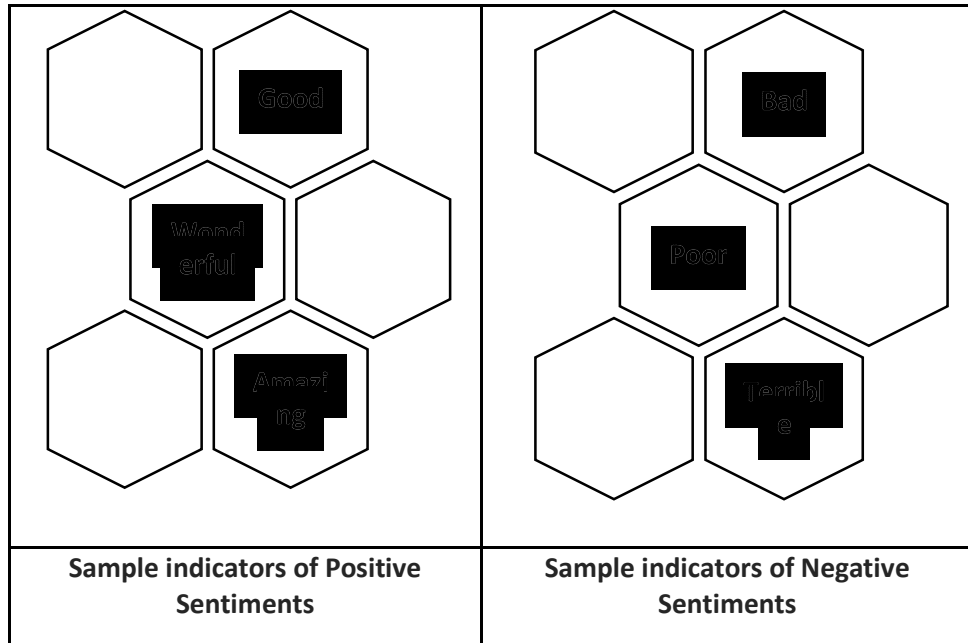
Figure 1

A popular method employed for determining sentiment is the use of a scaling system whereby words commonly associated with having a negative, neutral or positive sentiment with them are given an associated number on a -10 to +10 scale (Taboada, 2011).

This method makes it possible to adjust the sentiment of a given term relative to its environment (usually on the level of the sentence). When a piece of unstructured text is analysed using natural language processing, each concept in the specified environment is given a score based on the way sentiment words relate to the concept and its associated score.

The scaling system to determine the sentiment of a piece of text :

Negative	Neutral	Positive
-10 <---------------------	0 --------------------->	+10
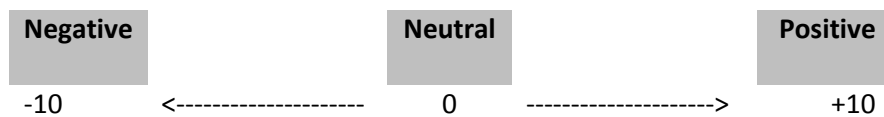
RapidMiner – a prominent open-source tool for text mining

RapidMiner is a user-friendly and open-source data science platform. The rich visual elements of the software accelerate the process of creating analytical models and makes it easy to get the results embedded in business operations. Defining analysis processes with RapidMiner Studio is done by drag and drop of operators, setting parameters and combining operators. The analysis processes can be produced from a large number of almost randomly nestable operators and finally be represented by flow-designs called process graphs. RapidMiner Studio offers the possibility of defining breakpoints and therefore inspecting virtually every intermediate result for better understanding and control of the analysis process. RapidMiner Studio contains more than 1500 operations for all tasks of professional data analysis, from data partitioning, to market-based analysis. It also supports text mining and automatic sentiment analysis (RapidMiner, 2016).
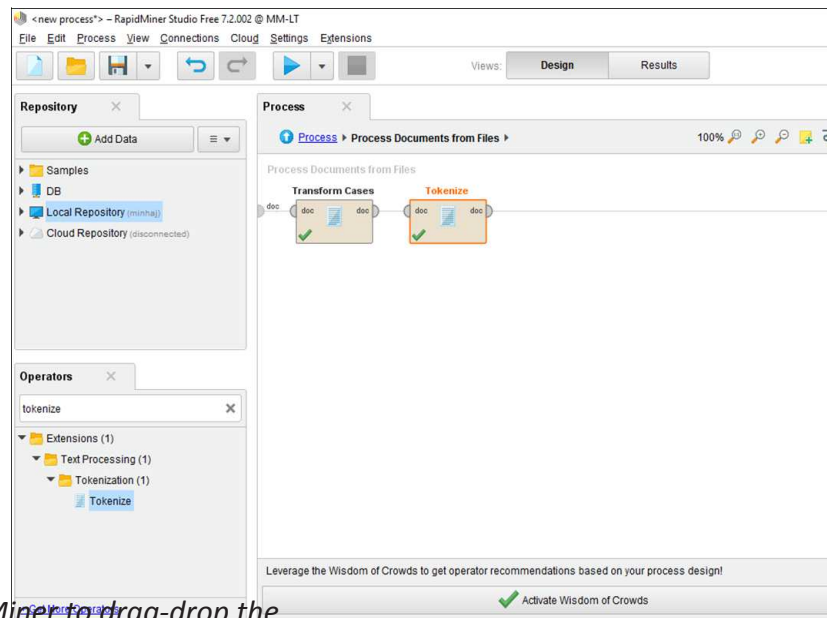
Figure 2 *: User friendly interface of RapidMiner to drag-drop the operators and create a mining process by connecting those operators*

Table 1

The operators and the process used in RapidMiner for Sentiment Analysis

RapidMiner Operator	Description
Operator to plug-in data into the RapidMiner Framework	The specific operator to be used to plug-in the data into the RapidMiner framework would depend on the type of data being analysed. Different types of data formats are supported by RapidMiner like database, excel file, twitter data etc. Also appropriate parameters have to be specified depending on the type of data operator used. For example, to analyse "Twitter Data", after adding the "Search Twitter" operator, specific parameters have to be configured pertaining to keyword to be searched, type of result (recent or popular), retweets to be ignored or not.
Nominal to Text	This operator changes the type of selected nominal attributes to text. It also maps all values of these attributes to corresponding string values.
Tokenize	This operator splits the text of a document into a sequence of tokens. There are several options for specifying the splitting points. You can either use all non-letter character, which is the default setting. This will result in tokens consisting of one single word, which is the most appropriate option before finally building the word vector or if you are going to build windows of tokens or you will probably split complete sentences, this is possible by setting the split mode to specify character and enter all splitting characters. The third option lets you define regular expressions and is the most flexible for very special cases. Each non-letter character is used as separator. As a result, each word in the text is represented by a single token.

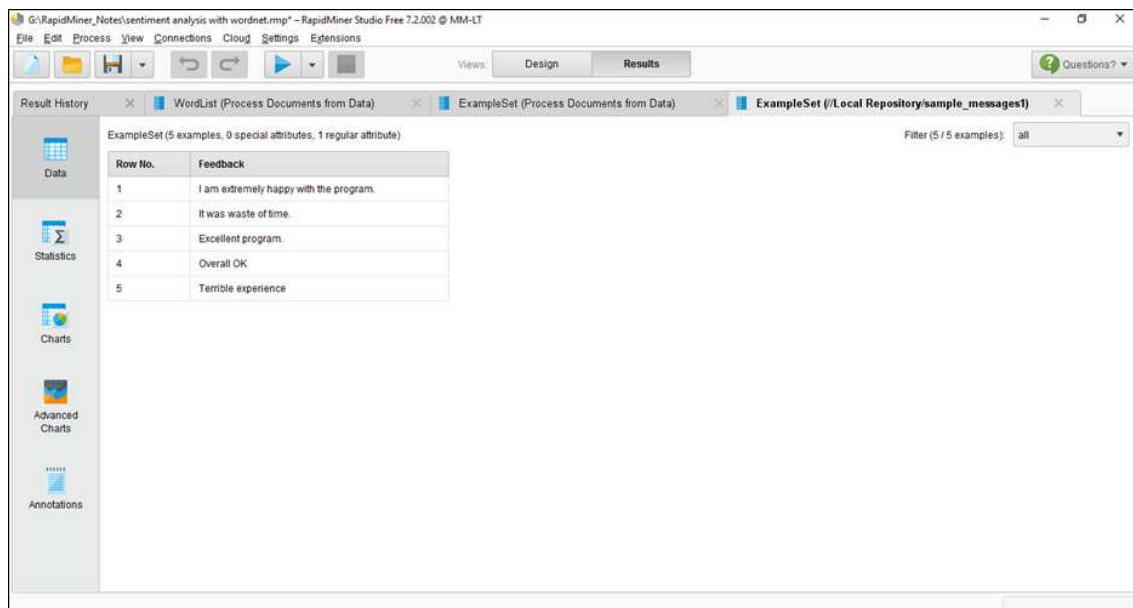| Transform case	This operator transforms all characters in a document to either lower case or upper case, respectively.
Filter Stop words	This operator filters English stop words from a document by removing every token which equals a stopword from the built-in stopword list. For this operator to work properly, every token should represent a single English word only.
Filter token (by length)	Filters tokens based on their length. This operator filters tokens based on their length (i.e. the number of characters they contain).
Open Wordnet dictionary	This operator facilitates use of the directory containing the Wordnet dictionary. WordNet® is a large lexical database of English. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. Synsets are interlinked by means of conceptual-semantic and lexical relations. (WordNet, 2016)
Extract Sentiment (English)	This operator uses a WordNet 3.0 and a SentiWordNet 3.0.0 database to extract sentiment of an input document. If the WordNet dictionary is not the same version as SentiWordNet, that is version 3.0, the operator will not work as expected. The sentiment value is in range [-1.0,1.0] where -1.0 means very negative and 1.0 means very positive. The sentiment value is added to the meta data of the document. WordNet and SentiWordNet are connected by Synset IDs
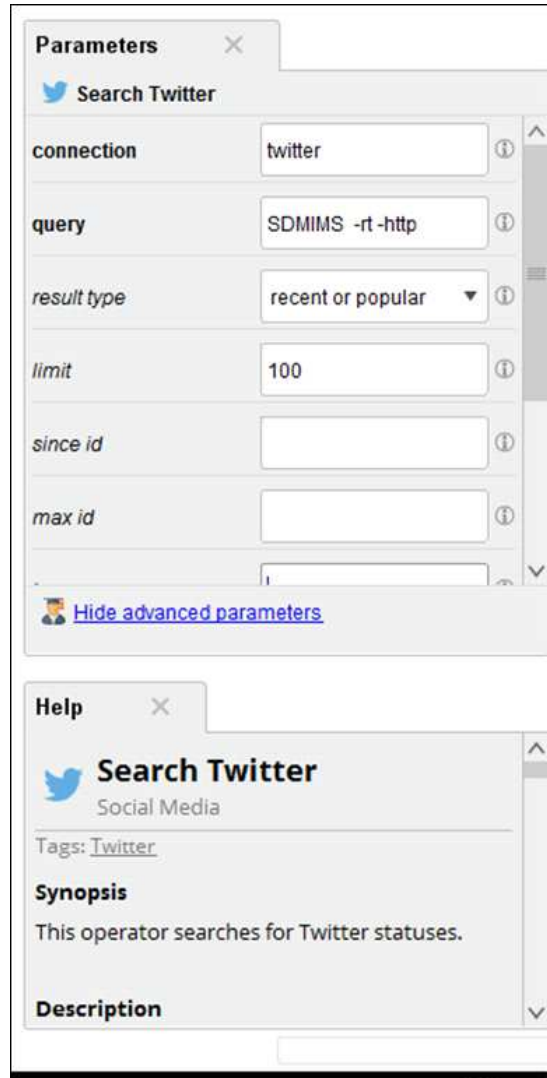
Figure 3 : *Parameters related to "Search Twitter" operator of RapidMiner*
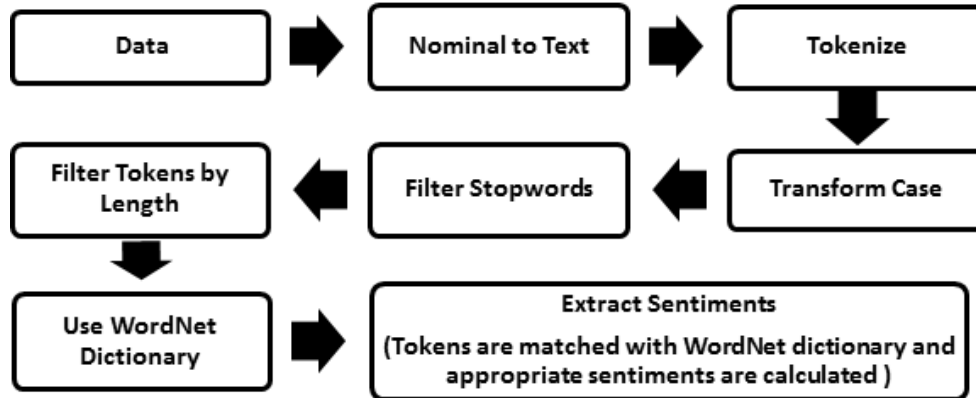
Figure 4 : *Process to Extract Sentiments from Text Sources in RapidMiner*

Sample input /output of the process to extract sentiments from text sources in RapidMiner

Figure 5 : Sample Input

Figure 6 : Sample Output with automatically created "Sentiment" column.

The sentiment analysis process explained in figure 4, involves plugging in the text corpus and processing the text - tokenising, transforming the case, filtering stop words (generic words like "a", "an", "the" "would" etc.,) and understanding the meaning of words using WordNet. Further with help of "Extract Sentiment" operator, the sentiment value of each text record is computed using SentiWorNet, which is internally computed as the average value of all the words' sentiments. The output of the process adds a new column "Sentiment" that provides a numeric value for the sentiment, negative sentiments are scored less than zero and positive sentiments are greater than zero (As depicted in Figure 6).

This general method of classifying pieces of texts into "Positive", "Negative" and "Neutral" can be used for variety of business applications like analysing customer feedback, evaluating product reviews etc.

References

(2016, 09 11). Retrieved from Oxford Dictionaries: https://en. oxforddictionaries.com/definition/sentiment_analysis

(2016, 09 12). Retrieved from RapidMiner: https://rapidminer. com/us/

(2016, 09 20). Retrieved from WordNet: https://wordnet. princeton .edu/

Liu, B. (2012). *Sentiment Analysis and Opinion Mining.* Morgan & Claypool Publishers.

Minhaj, M. (2016, 09 14). *Cloud Based BI*. Retrieved from SDMIMD: http://sdmimd.ac.in/SDMRCMS/articles/Cloud_Based_Business.pdf

Sentiment Analysis and Opinion Mining. (2016, 09 10). Retrieved from www.cs.uic.edu: https://www.cs.uic.edu/~liub/FBS/Sentiment Analysis-and-OpinionMining.pdf

Taboada, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 267-307.