**POLITECNICO DI MILANO**

Department of Electronics, Informatics and Bioengineering

M.Sc. programme in Computer Science and Engineering



Automated Analysis of Social Data using

Machine Learning Techniques

Advisors:

Prof. Letizia TANCA

Prof Maristella MATERA

Ing. Ricardo MEDANA

Master Thesis of:

Petar KORDA

Pavle VIDANOVIC

Academic Year 2017-2018

*We would like to express our sincere appreciation to our supervisors and our*

*families for all their support. We would also like to extend a special thank you to*

*Prof.ssa Maristella Matera for her time, guidance and understanding.*

**Abstract**

In today’s society everything is happening on the internet, in particular on social networks. Social networks play central role in everyday life of average person. So naturally, companies recognize the opportunity and try to make use of that by changing their business plans and focus to potential customers on social networks. Business is realized by company’s presence on web and producing a content that will take customer’s attention. In return users share their opinion about particular products by leaving comments on them and reacting on company’s posts.

Taking that into consideration, it is useful to have an automated way to check user reactions on products that company is offering. Also knowing types of people fallowing and leaving opinion on products can be turned into advantage for creating future business plans. For example, to predict which products can be attractive for specific user groups or to determine best time when to lunch products. We recognized the potential of that and that’s why we were eager to examine sentiment analysis tools and machine learning algorithms to achieve that goal.

In this thesis we have built automated for calculating sentiment of users who commented on specific company’s post along with intelligent spam filter. Sentiment analysis was done separately on text and on the emojis. For the evaluating text sentiment, we used open source API and for emojis we used table of evaluation for each emoji. Spam filter was designed using supervised machine learning techniques to determine spam, not just by searching URL patterns in comments, but also to determine the spam by checking text content. We have also built clustering module which uses unsupervised learning techniques on user data and visualizing characteristics for each discovered group. Finally, in last chapter is defined how previous models can be used together in an API to evaluate success of company’s posts.

**Sommario**

**Contents**

**Chapter 1**

**Introduction**

Analysis of social network content is difficult because conversation on social networks differs in many ways from normal conversation. Contents are rich with emojis, hashtags, mentions and spams which the one need to filter and process along with raw text to find the information behind it.

By analysing social networks data about certain products, brands or certain campaign can be very useful for companies and their business. In particular, companies can predict future trends, increase the profit and thus be in advantage over the competition.

Within previously described context, this project gives opportunity to user to analyse contents using sentiment analysis to determine sentiment of users on specific product, supervised machine learning algorithms within spam filter module to efficiently detect and remove spams from dataset and finally clustering module that discover user groups and their characteristics which can be used in making future predictions.

Sentiment analysis mentioned before represents 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 dictionary definition]. It is combined together with emoji sentiment evaluation in a way that sentiment analysis is done on the raw text using open source API while the emoji’s sentiment is evaluated using sentiment tables. Final sentiment is defined as specific combination of those two sentiments.

Spam filter module is equipped with two components. First part is done as text processor using regex expressions to detect links inside the text. Second part is trained Naïve Bayes machine learning model for detecting spams by checking the text context and represents more intelligent way of doing it.

Clustering module is created as unsupervised machine learning model, that taking users finds optimal number of user groups. It is also equipped with visualizing part that displays groups characteristics.

Finally, in the API part it is described how previous models can be used in one complete API that predicts successfulness of company’s posts on social networks.