# Frequently correlated terms in tweets

Data Mining Course Project Report

Davide Piva
2$^{\text{nd}}$ year CS Master Degree @ University of Trento
davide.piva-1@studenti.unitn.it

## 1 INTRODUCTION AND MOTIVATION

Twitter, as any other social network and blog, is a constant source of unstructured data that, if processed in the right way, can be leveraged to obtain valuable insights. Due to this reason, lots of companies have started to collect this huge amount of data in order to perform in-depth analysis; many of them have initiated to perform sentiment analysis to check customers satisfaction with respect to their products. According to upGrad [3], Big Data can help create pioneering breakthroughs for organizations that know how to use it correctly.

A challenging aspect in Big Data analysis is to extract valuable insights that can drive business strategies and decisions. In this context, Twitter plays a crucial role during the information retrieval phase: millions of people publish everyday short messages, hereafter tweets, that often contain their desires or demands. The aim of many companies is to exploit these tweets in order to catch common issues or demands to guide business strategies. These insights, according to upGrad [3], can be used to develop new products/services, improve marketing techniques, optimize customer service, improve employee productivity and find radical ways to expand brand outreach.

In addition, organizations aim also to intercept occasional issues or demands that are not constant over time but are very popular in a bounded period of time. Catching this kind of insights, permits companies to better address their business short-term decision-making phase. According to Ben Ridler [2], a business owner ability to effectively deal with customer complaints provides a great opportunity to turn dissatisfied customers into active promoters of the business.

Nowadays there are lots of datasets that group together even millions of tweets and many of them are easily and freely accessible by anyone through internet. Despite the fact that many of those datasets contain unusable data that force us to pre-process them, companies can elaborate those tweets to achieve their business goals.

## 2 RELATED WORK

## 3 PROBLEM STATEMENT

In order to achieve the above-mentioned goal (i.e., identify consistent topic in time in tweets), I have used a public available dataset that groups more than 300.000 tweets that contain the hashtag *#covid19* [1]. As many other datasets composed of unstructured data dumped from a social network such as Twitter, the dataset I used has lots of fields that characterize each tweet (e.g., publisher's username, location and account information, text of the message, etc.).

To identify consistent topics in time using those tweets, I only need the date of the publication of each of them and the relative text. In other words, in my working environment a tweet is defined as a tuple composed by the following fields:

- **date**: the date of the tweet publication, expressed as an integer value that represents the number of seconds that have elapsed since the midnight of 1$^{\text{st}}$ January 1970;
- **text**: the text of the tweet represented as a list of words. This list is obtained splitting the source text using the blank character " " as separator.

Therefore, in order to obtain the formal model of this problem, I define the following sets:

**TWEET** = the set of tweets, defined as TIMESTAMP × TEXT where TIMESTAMP $\subset \mathbb{N}_{\geq 0}$, TEXT $= \{x | x \in \Sigma^*\}$ and $\Sigma$ is the alphabet;

**THRESHOLD** = the set of all the possible thresholds to identify frequent terms in time, defined as a subset of $\mathbb{N}_{\geq 0}$

The aim is to model an utility function $f$ defined as

$$f : \text{TWEET} \times \text{THRESHOLD} \mapsto \text{TOPIC}$$

where TOPIC $= \{x | x \in \Sigma^*\}$ and $\Sigma$ is the alphabet. In other words, TOPIC is the set of all the possible trending topics over time.

## 4 SOLUTION

## 5 IMPLEMENTATION

## 6 DATASET

## 7 EXPERIMENTAL EVALUATION

## REFERENCES

[1] Covid19-Dataset [n.d.]. *Tweets with the hashtag covid19.* Retrieved December 17, 2020 from https://www.kaggle.com/gpreda/covid19-tweets

[2] Customer-Complaints [n.d.]. *Six Steps to Dealing with Customer Complaints.* Retrieved December 18, 2020 from https://www.eonetwork.org/octane-magazine/special-features/sixstepstodealingwithcustomercomplaints

[3] upGrad [n.d.]. *Benefits and Advantages of Big Data and Analytics in Business.* Retrieved December 17, 2020 from https://www.upgrad.com/blog/benefits-and-advantages-of-big-data-analytics-in-business/